ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

UNIVERSIDAD POLITÉCNICA DE MADRID

TESIS DE MÁSTER
MÁSTER EN INTELIGENCIA ARTIFICIAL

Clasificación Supervisada de las Neuronas de la Base de Datos NeuroMorpho

AUTOR : Patricia Maraver Abad
SUPERVISORES : Concha Bielza Lozoya
Pedro Larrañaga Mugica

Enero, 2015

Índice general

Índice de figuras **4**

Índice de tablas **7**

1. Introducción y Objetivos **9**
1.1. Objetivos 9
1.2. Las neuronas 11

2. Obtención de los Datos **14**
2.1. NeuroMorpho 14
2.2. L-Measure 17
2.3. Programa Java y Base de Datos mySQL 21

3. Tratamiento de los Datos **23**
3.1. Limpieza de datos erróneos 23
3.2. Datos no balanceados 24
3.3. Selección de atributos 26

4. Clasificación Supervisada **30**
4.1. Clasificadores 30
4.1.1. Clasificador bayesiano naïve 30
4.1.2. Árbol de clasificación C4.5 31
4.1.3. Clasificador IB1 32
4.1.4. Clasificador máquina de vectores soporte 33
4.2. Medidas de precisión 35

5. Resultados Obtenidos **37**
5.1. Clasificación de la especie 37
5.2. Clasificación del género 43
5.3. Clasificación de la edad 51
5.4. Clasificación por tipo de célula 61
5.5. Clasificación por región del cerebro 73
5.5.1. Clasificación general 73
5.5.2. Clasificación por región del neocórtex 79
5.6. Comparación de los algoritmos 83

6. Conclusiones **86**
6.1. Conclusiones 86
6.2. Trabajo Futuro 86

A. Atributos L-Measure **88**

B. Gráficas de resultados **90**
B.1. Clasificación de la especie 90
B.2. Clasificación del género 92
B.3. Clasificación de la edad 94
B.4. Clasificación del tipo de célula 96
B.5. Clasificación por región del cerebro 98
B.5.1. Clasificación general 98
B.5.2. Clasificación por región del neocórtex 100

C. Bibliografía **103**

Índice de figuras

1.1. Estructura de una neurona 13

2.1. Captura de NeuroMorpho: acceso por especie 15
2.2. Captura de NeuroMorpho: acceso por metadatos 15
2.3. Neuronas: capturas de la animación 3D de la base de datos de NeuroMorpho. 16
2.4. Conceptos utilizados por el software L-Measure 17
2.5. Atributos de L-Measure (I) 18
2.5. Atributos de L-Measure (II) 19
2.5. Atributos de L-Measure (III) 20
2.5. Atributos de L-Measure (IV) 21

3.1. Matriz de confusión de la clasificación de neuronas con datos no balanceados . 24
3.2. Matriz de confusión de la clasificación de neuronas con datos balanceados . . 26

4.1. Estructura de red bayesiana naïve 31
4.2. Gráfica del efecto de usar una Gausiana frente al método kernel para estimar la densidad de una variable continua (John y Langley (1995)) 31
4.3. Ejemplos de distintos hiperplanos de margen máximo (imágenes de *http://www.support-vector-machines.org*) 34
4.4. Matriz de confusión de un clasificador binario 35

5.1. Distribución de las neuronas por especie y tipo de neurona 38
5.2. Matrices de confusión del clasificador IB1 por especies para 12 clases 40
5.3. Curva ROC 41
5.4. Matrices de confusión del clasificador bayesiano naïve por especies para los terminales axónicos 43
5.5. Varianza de cada clasificador para los 10 modelos con selección wrapper de clasificación de la especie para 70 instancias por clase 43
5.6. Neuronas de NeuroMorpho por género 45
5.7. Matrices de confusión del clasificador SVM del género para la especie humana 46
5.8. Varianza de cada clasificador para los 10 modelos de clasificación del género para la especie humana 47
5.9. Varianza de cada clasificador para los 10 modelos de clasificación del género para la especie de las ratas 47
5.10. Árbol de clasificación C4.5 del género para la especie de las ratas 48
5.11. Matrices de confusión del árbol de clasificación C4.5 del género para la especie de las ratas 48

5.12. Matrices de confusión del clasificador SVM del género para las células principales de la especie de las ratas 49
5.13. Varianza de cada clasificador para los 10 modelos de clasificación del género para todas las especies 50
5.14. Matrices de confusión del clasificador IB1 del género para todas las especies . 50
5.15. Burke y Barnes (2006): Evolución morfológica de las neuronas con la edad. (a) Neurona humana; (b) neurona de rata 52
5.16. Neuronas de NeuroMorpho por edad (vertical), especie y tipo (horizontal) . . 53
5.17. Matrices de confusión del clasificador bayesiano naïve por edades para la especie humana 54
5.18. Modelo bayesiano naïve de clasificación por edades para las células principales de la especia humana. Se muestran los 5 primeros valores de cada atributo junto con sus probabilidades y frecuencias 55
5.19. Varianza de cada clasificador para los 10 modelos de clasificación por edades para la especie humana 55
5.20. Matrices de confusión del clasificador bayesiano naïve ((a) y (b)) y del clasificador IB1 ((c) y (d)) para clasificar las edades de la especie rata 56
5.21. Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie rata 57
5.22. Matrices de confusión del clasificador SVM de clasificación de la edad para la especie ratón 58
5.23. Matrices de confusión del clasificador SVM de clasificación de la edad para las interneuronas de la especie ratón 58
5.24. Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie ratón para los distintos grupos de neuronas 59
5.25. Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie drosophila para los distintos grupos de neuronas 59
5.26. Matrices de confusión del clasificador bayesiano naïve por edades para la especie drosophila 60
5.27. Matrices de confusión del clasificador SVM de la edad para los terminales axónicos de la especie drosophila 61
5.28. Distribución del tipo de célula en NeuroMorpho 62
5.29. Tipos de neuronas: neuronas excitatorias en negro, neuronas inhibitorias en rosa. Ba, *basket cells*; Bi,*bipolar cell*; Ch, *Chandelier cell*; DB, *double bouquet cell*; HC: *horizontal cell of Cajal o Cajal-Retzius*; M, *Martinotti cell*; N, *neurogliaform*; P, *pyramidal neurons*; SS, *spiny stellate cells*; I_i, diferentes tipos de interneuronas. Imagen obtenida de Nieuwenhuys *et al.* (2007) 64
5.30. Matrices de confusión del clasificador IB1 por tipo de célula 65
5.31. Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los criterios de Ramón y Cajal 67
5.32. Matrices de confusión del clasificador bayesiano naïve para la clasificación por tipo de célula atendiendo a los criterios de los investigadores 68
5.33. Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los criterios de forma 69
5.34. Matrices de confusión del clasificador bayesiano naïve para la clasificación por tipo de célula atendiendo a los criterios moleculares 70

5.35. Matrices de confusión del clasificador bayesiano naïve la clasificación por tipo de célula atendiendo a los criterios de funcionalidad 71
5.36. Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los axones 72
5.37. Neuronas piramidales de diferentes áreas corticales. Imagen obtenida de Spruston (2008) 73
5.38. Región del cerebro 74
5.39. Curva ROC 76
5.40. Matrices de confusión del clasificador bayesiano naïve para la clasificación de la región del cerebro para las interneuronas 78
5.41. Distribución de las neuronas de NeuroMorpho por región del neocórtex 79
5.42. Matrices de confusión del clasificador IB1 para la clasificación de las células principales por región del neocórtex 81
5.43. Matrices de confusión del clasificador IB1 para la clasificación de las interneuronas por región del neocórtex 83

B.1. Clasificación de la especie para el clasificador bayesiano naïve 90
B.2. Clasificación de la especie para el clasificador j48 91
B.3. Clasificación de la especie para el clasificador IB1 91
B.4. Clasificación de la especie para el clasificador SVM 92
B.5. Clasificación por género para el clasificador bayesiano naïve 92
B.6. Clasificación por género para el clasificador j48 93
B.7. Clasificación por género para el clasificador IB1 93
B.8. Clasificación por género para el clasificador SVM 94
B.9. Clasificación por edad para el clasificador bayesiano naïve 94
B.10. Clasificación por edad para el clasificador j48 95
B.11. Clasificación por edad para el clasificador IB1 95
B.12. Clasificación por edad para el clasificador SVM 96
B.13. Clasificación por tipo de célula para el clasificador bayesiano naïve 96
B.14. Clasificación tipo de célula para el clasificador j48 97
B.15. Clasificación por tipo de célula para el clasificador IB1 97
B.16. Clasificación por tipo de célula para el clasificador SVM 98
B.17. Clasificación por región del cerebro para el clasificador bayesiano naïve 98
B.18. Clasificación por región del cerebro para el clasificador j48 99
B.19. Clasificación por región del cerebro para el clasificador IB1 99
B.20. Clasificación por región del cerebro para el clasificador SVM 100
B.21. Clasificación del neocórtex para el clasificador bayesiano naïve 100
B.22. Clasificación del neocórtex para el clasificador j48 101
B.23. Clasificación del neocórtex para el clasificador IB1 101
B.24. Clasificación del neocórtex para el clasificador SVM 102

Índice de tablas

1.1. Resumen del estado del arte para la clasificación de neuronas 12

2.1. Ejemplo del archivo swc para la neurona del gusano c-elegans de la figura 2.3 (c). Se muestran 6 compartimentos de los 53 totales 16

3.1. Neuronas extraídas de NeuroMorpho 23

5.1. Tabla de resultados del clasificador IB1 variando el número de instancias y clases 39
5.2. Atributos obtenidos con selección wrapper para el modelo IB1 de clasificación por especies para 12 clases 42
5.3. Atributos obtenidos con selección de atributos CFS para el modelo SVM de clasificación del género para la especie humana 46
5.4. Atributos obtenidos con selección de atributos wrapper para el árbol de clasificación C4.5 del género para la especie de las ratas 48
5.5. Atributos obtenidos con selección de atributos wrapper para la clasificación SVM del género para la especie de las ratas 49
5.6. Atributos obtenidos con selección de atributos wrapper para la clasificación IB1 por género de todas las especies 51
5.7. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve de clasificación por edades para la especie humana 53
5.8. Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación de la edad de la especie rata 57
5.9. Atributos obtenidos con selección de atributos wrapper para el modelo SVM de clasificación de la edad para la especie ratón 58
5.10. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve de clasificación por edades para la especie drosophila 60
5.11. Atributos obtenidos con selección de atributos wrapper para el modelo SVM de clasificación de la edad para los terminales axónicos de la especie drosophila 61
5.12. Atributos obtenidos con selección de atributos wrapper para el modelo IB1 para la clasificación del tipo de célula 66
5.13. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de Ramón y Cajal 67
5.14. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de los investigadores 69

5.15. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de forma 69
5.16. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios moleculares 70
5.17. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de funcionalidad 71
5.18. Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los axones . . 72
5.19. Tabla de resultados del clasificador IB1 variando el número de instancias y clases 75
5.20. Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación por regiones del cerebro 77
5.21. Regiones del cerebro por especies para las interneuronas 78
5.22. Regiones del cerebro por especies para los terminales axónicos 79
5.23. Atributos obtenidos con selección de atributos wrapper para el clasificador IB1 de las células principales por región del neocórtex 82
5.24. Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación de interneuronas por región del neocórtex 83
5.25. Mejora obtenida con la selección de atributos en los distintos clasificadores . . 84
5.26. Número medio de atributos seleccionados en los distintos clasificadores 84

A.1. Atributos del software L-Measure 89

Capítulo 1

Introducción y Objetivos

1.1. Objetivos

A principios de los años 80 se realizaron grandes avances en los estudios morfológicos cerebrales centrados, especialmente, en el tamaño, forma, conexiones neuronales y número de neuronas de las distintas zonas del cerebro. En esa época empezaban también los estudios con microscopio que utilizaban la técnica de Golgi, la cual revela la morfología de la neurona completa en tres dimensiones. Gracias a esta técnica y a la microscopía digital utilizada en los últimos años se han publicado investigaciones sobre las diferencias morfológicas entre las neuronas de distintas especies, géneros, tipo de célula, región del cerebro y edad.

Partimos de una colección de neuronas digitalizadas que descargaremos de la mayor base de datos libre y accesible vía web que existe actualmente llamada NeuroMorpho (Ascoli *et al.* (2007)) y ubicada en *http://neuromorpho.org*. A partir de los atributos que extraeremos de las células con el software L-Measure clasificaremos las distintas neuronas por especies, género, tipo de célula, región del cerebro y edad utilizando los algoritmos de aprendizaje automático disponibles en el software Weka. Por último estudiaremos los resultados obtenidos.

En el capítulo de resultados obtenidos se describen los datos presentados por los distintos investigadores que han realizado los estudios manualmente, tratando las neuronas una a una y los compararemos con los que hemos obtenido computacionalmente. Veremos las diferencias y similitudes, y podremos verificar la robustez de nuestros resultados. Gracias a la capacidad actual de los ordenadores y a los avances en inteligencia artificial descubriremos atributos para diferenciar clases que no se conocían por las limitaciones humanas, además de poder ratificar aquellos que ya se utilizan.

Existen investigaciones previas basadas en técnicas de aprendizaje no supervisado para obtener grupos o *clusters* de neuronas respecto del tipo de neurona (células piramidales, multipolares, en cesta ...). Tsiola *et al.* (2003) agrupan las neuronas del córtex primario visual en ratones utilizando clustering multidimensional. Realizan la extracción de la morfología de la neurona con técnicas de Golgi y reducen el número de características utilizando el análisis de componentes principales. Como resultado obtienen cinco grupos que se distinguen por la extensión de sus neuronas. En las publicaciones posteriores a 2006 se utilizan las reconstrucciones y las características neuronales que ofrece la base de datos NeuroMorpho. Chunwen *et al.* (2011) aplican clustering jerárquico sobre sesenta neuronas descargadas aleatoriamente. Al

igual que Tsiola *et al.* (2003), utilizan el análisis de componentes principales pero con células de distintas especies. Obtienen cuatro categorías que verifican observando las reconstrucciones neuronales que ofrece NeuroMorpho y concluyen que las neuronas del mismo tipo son parecidas entre especies. En cambio McGarry *et al.* (2010) obtienen tres categorías aplicando clústering jerárquico a las interneuronas del córtex cerebral. Zawadzki *et al.* (2012) utilizan el análisis de componentes principales para reducir los atributos a un espacio bidimensional. Obtienen la probabilidad de pertenencia a cada grupo de células con un estimador kernel de densidad bivariante clasificando las de menor probabilidad como valores atípicos y los arquetipos como células más probables para el tipo tratado. Para el estudio utilizaron cinco mil neuronas lo que supone un aumento considerable respecto de los estudios anteriores que no superan las 100 neuronas. En la publicación de Yu *et al.* (2012), mediante el algoritmo de esperanza-maximización, dividen las neuronas en seis clústers que verifican con una red bayesiana naïve y por último predicen el crecimiento neuronal pero, a diferencia de los anteriores, extraen setenta y ocho atributos con el software Neuron. Guerra *et al.* (2013) proponen un algoritmo de aprendizaje semi-supervisado de clustering probabilístico para la clasificación por tipos de 241 interneuronas utilizando nueve atributos obtenidos con el software Neurolucida. Por otro lado, Wong *et al.* (2002) estudian el patrón de los axones en el mapa olfativo de la mosca drosophila utilizando clustering jerárquico.

Cauli *et al.* (2000), Karagiannis *et al.* (2009), Druckmann *et al.* (2013) y McGarry *et al.* (2010) aplican clustering para clasificar los tipos de neuronas a nivel molecular o electrofisiológico. McGarry *et al.* (2010) confirman con el estudio molecular el realizado mediante análisis morfológico al obtener los tres mismos grupos.

En aprendizaje supervisado Fengqing y Jie (2012) proponen utilizar el algoritmo de máquinas de vectores soporte donde algunos de los atributos o características son además de los obtenidos con el software L-Measure medidas fractales que obtienen de las células. Entrenando el algoritmo con apenas cuarenta y cuatro neuronas y testeándolo con veinte. En cambio Xianhua (2011) aplica análisis discriminante bayesiano a partir de cinco atributos que ha considerado relevantes de los 43 obtenidos con el software L-Measure y distingue células piramidales por especies: mono y rata. Guerra *et al.* (2011) diferencian neuronas piramidales e interneuronas del neocórtex del ratón y comparan el clustering jerárquico con los algoritmos de clasificación supervisada: red bayesiana naïve, árbol de clasificación C4.5, k-vecinos más cercanos k-nn, perceptrón multicapa y regresión logística, utilizando tanto análisis de componentes principales como selección de atributos. Por último Marin *et al.* (2002) han obtenido observaciones similares a Wong *et al.* (2002) estudiando el patrón de los axones de la mosca drosophila utilizando análisis lineal discriminante. Presentamos un resumen del estado del arte en la tabla 1.1.

En nuestro trabajo utilizaremos el conjunto completo de variables que proporciona el software L-Measure (capítulo 2 sección 2.2), no sólo las propuestas por el grupo de investigación de NeuroMorpho. Reduciremos el espacio de variables con técnicas multivariantes (descritas en la sección 3.3 del capítulo 3) obteniendo atributos relevantes que no aparecen en dicha base de datos y que las investigaciones han pasado por alto. Descargaremos las 11335 neuronas que contiene la base de datos en la versión 5.7 publicada el 30 de mayo de 2014 (capítulo 2 sección 2.1), permitiéndonos realizar estudios más detallados. Clasificaremos no sólo por tipo de célula sino por especie, género, región del cerebro y edad (capítulo 5) utilizando todas las neuronas disponibles, asumiendo a diferencia de otras publicaciones, que las neuronas son

diferenciables por cada clase considerada, independientemente del resto. Por ejemplo, serán clasificables por especie sin tener en cuenta la edad, el género, la zona del cerebro y el tipo de célula.

Utilizaremos distintas técnicas de aprendizaje automático supervisado (descritas en el capítulo 4): una red bayesiana, un árbol de decisión, un algoritmo basado en instancias y el algoritmo máquina de vectores soporte. Haremos una comparativa entre los distintos algoritmos y sus resultados, y extraeremos conclusiones sobre la precisión de los resultados y el rendimiento computacional tanto en tiempo como en memoria de los mismos (sección 5.6 del capítulo 4).

A continuación describimos los conceptos fundamentales de las neuronas que serán necesarios para entender las agrupaciones que hemos realizado para la posterior clasificación.

1.2. Las neuronas

Las neuronas son las células del sistema nervioso que está formado por el cerebro, la médula espinal, los sistemas sensoriales periféricos y el sistema entérico (intestinal). Están formadas por un cuerpo celular central llamado soma donde se encuentra el núcleo de la célula. En general, las neuronas tienen un polo de entrada que consiste en una ramificación con forma de árbol con múltiples extensiones llamadas dendritas y que salen directamente del cuerpo celular en los vertebrados. En cambio en los invertebrados salen normalmente desde el polo de salida conocido como axón, aunque existen algunas dendritas que también funcionan como salida.

El botón sináptico o terminal axónico es la estructura final del axón encargado de transmitir la información entre neuronas. Recibe el impulso eléctrico que oscila entre 50 y 70 milivoltios que transforma en una señal química conocida como neurotransmisor y que comunica con las dendritas de la neurona siguiente. Algunos laboratorios estudian el axón como una parte separada del resto de la neurona por su importancia. Podemos ver en la figura 1.1 un ejemplo de la estructura de una neurona.

	Referencia	Algoritmo	Especie	Región del cerebro	Obtención de las neuronas	Nº neuronas	Obtención de los atributos	Nº atributos iniciales	Reducción de atributos	Objetivo
Aprendizaje no supervisado	Tsiola *et al.* (2003)	Clústering jerárquico	Ratón	Capa V del córtex primario visual	Impregnaciones de Golgi, inyecciones de biocitina y DiOlistics	158	Neurolucida	33 morfológicos	Análisis de componentes principales	Tipos de células
	Chunwen *et al.* (2011)	Clústering jerárquico	Varias	Varias	NeuroMorpho	60	L-Measure	20 morfológicos	Análisis de componentes principales	Tipos de células
	McGarry *et al.* (2010)	Clústering jerárquico	Ratón	Córtex frontal, visual y somatosensorial	Biocitina	59	Grabaciones clamp y Neurolucida	67 morfológicos 19 electrofisiológicos	Análisis de componentes principales	Tipos de interneuronas
	Zawadzki *et al.* (2012)	Estimador kernel de densidad	Varias	Varias	NeuroMorpho	5000	L-Measure	20 morfológicos	Análisis de componentes principales	Arquetipos vs valores atípicos
	Yu *et al.* (2012)	Clustering probabilístico	Humana	Neocórtex	NeuroMorpho	1907	Neuron	78 morfológicos	-	Predicción del crecimiento neuronal
	Wong *et al.* (2002)	Clústering jerárquico	Drosophila	Protocerebro	Impregnaciones de Golgi	36	Neurolucida	21 morfológicos	Análisis de componentes principales	Mapa olfativo
	Cauli *et al.* (2000)	Clústering jerárquico	Rata	Neocórtex	Óptica Nomarski y videomicroscopía infrarroja	60	Grabaciones clamp	14 electrofisiológicos	-	Tipos de interneuronas
	Karagiannis *et al.* (2009)	Clústering jerárquico	Rata	Neocórtex	Biocitina	200	ImagePro 5.1, Grabaciones clamp, scRT-PCR	12 morfológicos, 32 electrofisiológicos, 10 moleculares	-	Tipos de interneuronas
	Druckmann *et al.* (2013)	Clústering jerárquico	Rata	Neocórtex	Grabaciones intracelulares in vitro	466	-	38 electrofisiológicos	Análisis de componentes principales, selección de atributos exhaustiva, selección de atributos ramificación y poda	Tipos de interneuronas
	Guerra *et al.* (2011)	Clústering jerárquico	Ratón	Neocórtex	Biocitina	327	Neurolucida	65 morfológicos	Análisis de componentes principales, selección de atributos (filtro)	Diferencia entre neuronas piramidales e interneuronas
Aprendizaje semi-supervisado	Guerra *et al.* (2013)	Clústering probabilístico	Varias	Neocórtex	NeuroMorpho	241	Neurolucida	9 morfológicos	Análisis de componentes principales, selección de atributos	Tipos de interneuronas
Aprendizaje supervisado	Fengqing y Jie (2012)	Máquinas de vectores soporte	Varias	Varias	NeuroMorpho	64	L-Measure, medidas fractales	17 morfológicos y fractales	-	Tipos de células
	Xianhua (2011)	Análisis discriminante bayesiano	Varias	Varias	NeuroMorpho, otros	364	L-Measure	43 morfológicos	Selección subjetiva	Tipos de células, especies (ratón y mono)
	Guerra *et al.* (2011)	Red bayesiana naïve, árbol C4.5, K-vecinos más cercanos knn, perceptrón multicapa, regresión logística	Ratón	Neocórtex	Biocitina	327	Neurolucida	65 morfológicos	Análisis de componentes principales, selección de atributos (filtro, envoltura)	Diferencia entre neuronas piramidales e interneuronas
	Marin *et al.* (2002)	Análisis lineal discriminante	Drosophila	Protocerebrum	Mosaico genético	1300	Neurolucida	37 morfológicos	selección de atributos exhaustiva	Mapa olfativo

Tabla 1.1: Resumen del estado del arte para la clasificación de neuronas

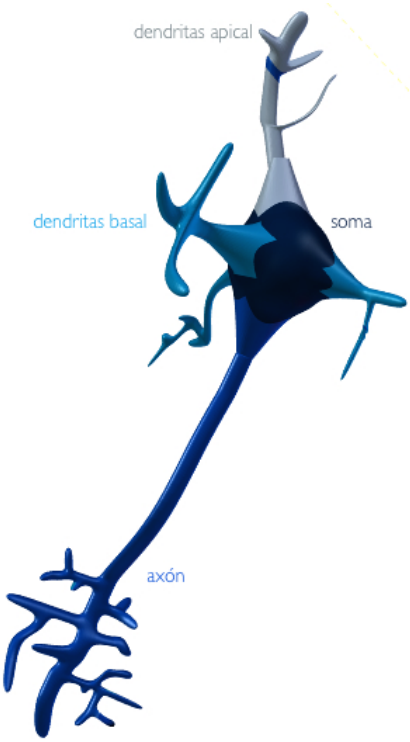


Figura 1.1: Estructura de una neurona

En NeuroMorpho el primer nivel en la clasificación por tipo de neurona diferencia entre interneuronas y células principales. Las redes corticales contienen estos dos elementos básicos (Gulyas *et al.* (1998), Buzsáki y Chrobak (1995)). La mayoría de las neuronas, entre el 80 % y el 90 % son células principales excitatorias que tienen proyecciones axonales fuera del área donde se encuentra el núcleo de la célula y las dendritas, conectando con otras zonas corticales. Las células inhibitorias o *interneuronas* representan entre el 10 % y el 20 % del total de las neuronas, descritas en profundidad en Freund y Buzsáki (1996), se encuentran en todas las áreas y capas del córtex cerebral y poseen axones locales. Además se diferencian de las anteriores en las propiedades fisiológicas.

Capítulo 2

Obtención de los Datos

Hemos descargado la totalidad de las neuronas de NeuroMorpho. Una vez procesadas con el software L-Measure para la extracción de características de cada neurona se han guardado los resultados en una base de datos local MySQL. Todo este proceso se ha automatizado con una aplicación Java para permitir la inclusión de nuevas neuronas en el futuro.

2.1. NeuroMorpho

NeuroMorpho es un archivo digital de acceso gratuito vía *web* (*http://neuromorpho.org*) que contiene la reconstrucción de la morfología neuronal. Encontramos una descripción completa en Ascoli *et al.* (2007) y posteriormente ampliada en Halavi *et al.* (2008). Actualmente está formado por 11335 neuronas reconstruidas digitalmente, todas ellas asociadas con su correspondiente publicación. Para la creación y actualización de la base de datos contribuyen alrededor de 100 laboratorios de todas las partes del mundo.
Se permite el acceso a las neuronas según varios criterios (figura 2.1): la región del cerebro, la especie animal, el tipo de célula, el laboratorio. Además podemos realizar búsquedas basadas en metadatos como la edad del animal (figura 2.2).

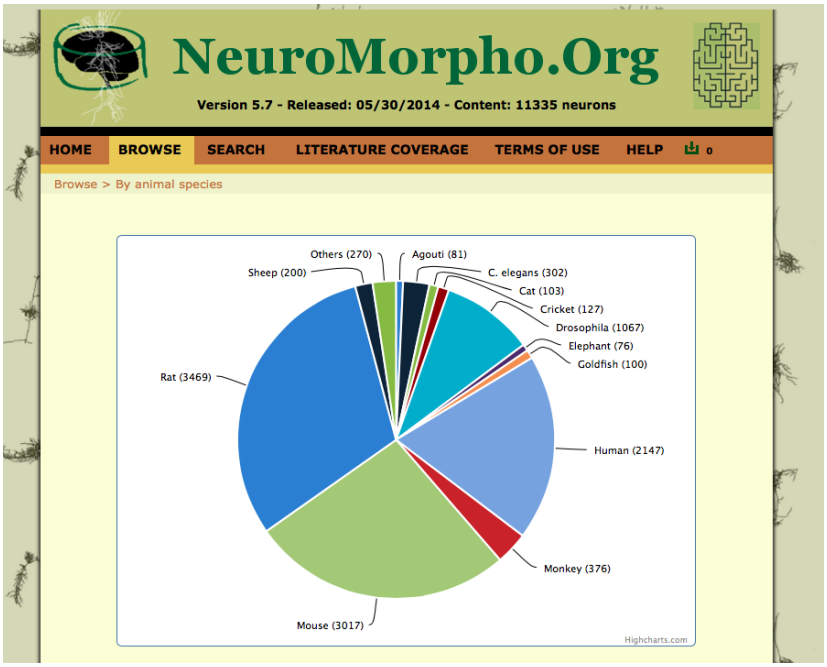


Figura 2.1: Captura de NeuroMorpho: acceso por especie

Figura 2.2: Captura de NeuroMorpho: acceso por metadatos

En la página *web* se visualiza la reconstrucción en 2D y 3D de la neurona. Tenemos un ejemplo de distintas neuronas en la figura 2.3.

(a) mosca drosophila (b) humana (c) gusano c. elegans

Figura 2.3: Neuronas: capturas de la animación 3D de la base de datos de NeuroMorpho.

La estructura tridimensional de una neurona se puede representar como un archivo ascii de formato *swc* propuesto originalmente por Cannon *et al.* (1998). En dicho archivo cada línea contiene siete campos de datos codificados para cada compartimento de la neurona (tabla 2.1): un índice que lo identifica, el tipo (sin definir, soma, axón, dendrita basal o dendrita apical), sus coordenadas (x, y, z), el radio, y el índice del compartimento padre.

Índice	Tipo	Coordenada x	Coordenada y	Coordenada z	Radio	Compartimento padre
1	1 (soma)	0.0	0.0	0.0	2.051	-1 (origen)
2	1 (soma)	0.0	1.9	0.77	2.051	1
3	1 (soma)	0.0	-1.9	-0.77	2.051	1
4	2 (axón)	-1.4	0.42	0.13	0.283	1
5	2 (axón)	-3.12	0.09	0.93	0.3	4
6	2 (axón)	-4.48	-0.93	3.45	0.287	5

Tabla 2.1: Ejemplo del archivo swc para la neurona del gusano c-elegans de la figura 2.3 (c). Se muestran 6 compartimentos de los 53 totales

Cada compartimento padre que sea el origen de la célula contendrá el valor -1 que generalmente forma parte del soma. Todos los demás compartimentos tendrán sólo un padre. Aunque pueden darse múltiples bifurcaciones en un solo punto, cada padre tendrá un índice inferior al compartimento hijo. Los bucles y las ramas aisladas se excluyen. El soma se codifica como una esfera si se trata de una línea, pero si contiene varias, como ocurre en las neuronas piramidales, se representará como un conjunto de cilindros.

NeuroMorpho asocia a cada neurona unos valores morfológicos, que a diferencia de los ficheros swc, no son descargables. Esos valores se han obtenido utilizando el software L-Measure. Nosotros hemos utilizado el mismo software para extraer y ampliar las características

que se exponen en la *web*, como describimos a continuación.

2.2. L-Measure

Scorcioni *et al.* (2008) han creado L-Measure, una herramienta software de libre distribución que caracteriza la morfología neuronal mediante un conjunto de métricas que extrae a partir de archivos digitales de neuronas swc. Está disponible a través de la *web* en la dirección *http://cng.gmu.edu:8080/Lm/*.

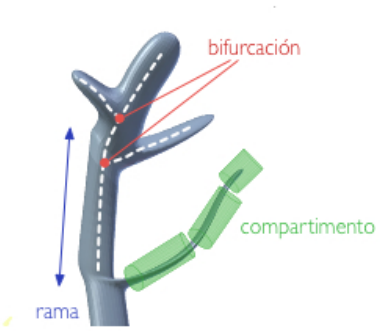
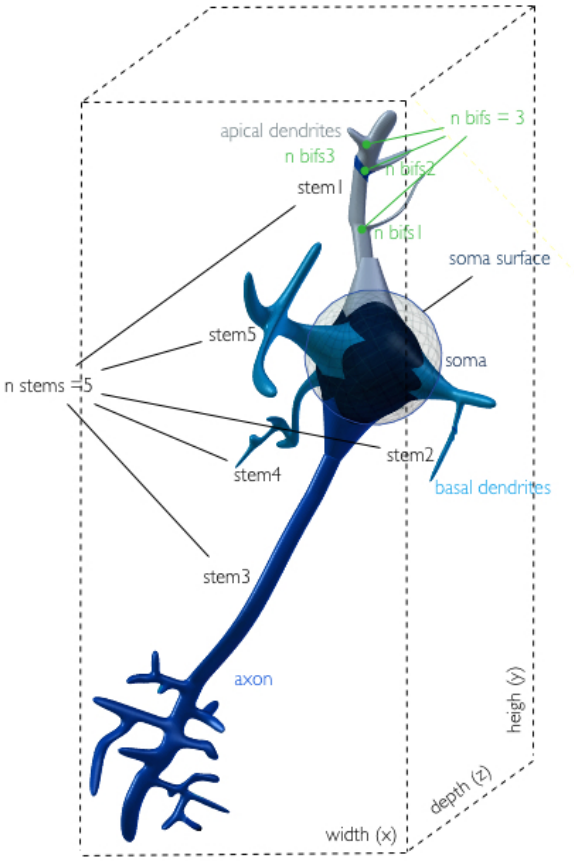


Figura 2.4: Conceptos utilizados por el software L-Measure

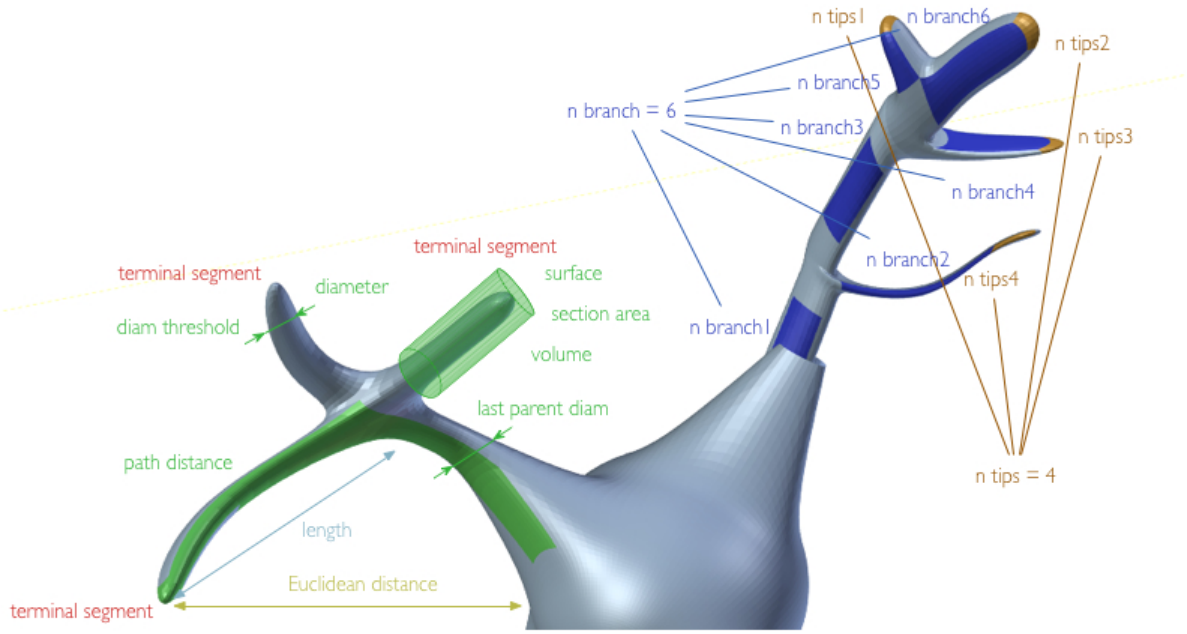
Los conceptos fundamentales sobre la neurona que maneja el software son los siguientes. *Los compartimentos* son segmentos representados como cilindros con diámetro, cada línea del archivo swc representa un compartimento definido por las coordenadas de los puntos extremos (x, y, z). *Las ramas* están formadas por uno o más compartimentos, una rama empieza y termina en una bifurcación. *Las bifurcaciones* son los puntos donde una rama se divide en dos o más ramas como vemos en la figura 2.4.

Este software es capaz de calcular 43 medidas características de las neuronas. De cada una de ellas se obtiene la suma, la media, el máximo, el mínimo y la desviación estándar siendo en total 215 atributos, ilustrados en la figura 2.5 y resumidos en la tabla A.1 del anexo. Podemos ampliar la información de cada medida en *http://farsight-toolkit.org/wiki/L_Measure_functions*.



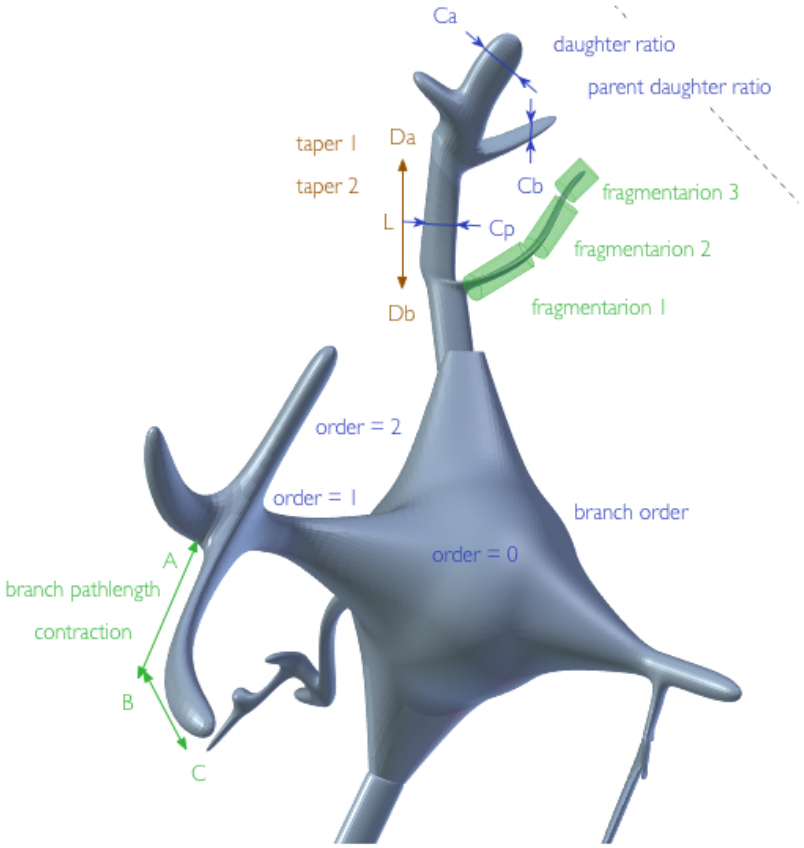
(a) Soma surface, n stems, n bifs, width, height, depth, type (soma=1, axon=2, basal dendrites=3, apical dendrites=4)

Figura 2.5: Atributos de L-Measure (I)



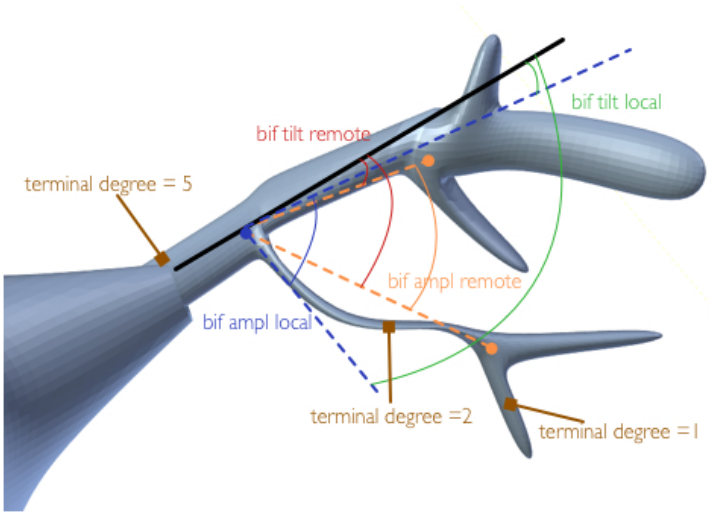
(b) N branch, n tips, diameter of the compartment, diameter pow ($diameter^{1.5}$), length of the compartment, surface of the compartment (area of cylinder of the compartment: $2 * \pi * r * h$), section area of the compartment (area of cylinder of the compartment: $\pi * r^2$), volume of the compartment (volume of cylinder of the compartment: $\pi * r^2 * h$), Euclidean distance of a compartment with respect to soma ($\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$), path distance (summation of the individual compartment lengths that form a branch), terminal segment (1 for all the compartments in the terminal branch), last parent diam (the diameter of last bifurcation compartment before the terminal tips), diam threshold (diameter of first compartment after the last bifurcation leading to a terminal tip), hillman threshold ($0.5 * ParentDiameter + 0.25 * DaughterDiameter_1 + 0.25 * DaughterDiameter_2$)

Figura 2.5: Atributos de L-Measure (II)



(c) Branch order, taper 1 ($\frac{D_a - D_b}{L}$), taper 2 ($\frac{D_a - D_b}{D_a}$), branch pathlength ($AB + BC$), contraction ($\frac{AC}{AB+BC}$), fragmentation (number of compartments that constitute a branch between two bifurcation points or between a bifurcation point and a terminal tip), daughter ratio ($\frac{C_a}{C_b}$), parent daughter ratio ($\frac{C_a}{C_p}$, $\frac{C_b}{C_p}$), rall power ($D_p = D_a^n + D_b^n$, the final rall value (n) is the idealistic n value that can propagate the signal transmission without loss from the starting point to the terminal point in a cable model assumption), Pk ($\frac{D_a^n + D_b^n}{D_p^n}$), Pk classic ($\frac{D_a^1.5 + D_b^{1.5}}{D_p^{1.5}}$), Pk 2 ($\frac{D_a^2 + D_b^2}{D_p^2}$)

Figura 2.5: Atributos de L-Measure (III)



(d) Terminal degree (total number of tips that each compartment will terminate into), bif ampl local (the angle between the first two compartments in degree), bif ampl remote (the angle between two bifurcation points or between bifurcation point and terminal point or between two terminal points), bif tilt local (the angle between the previous compartment of bifurcating father and the two daughter compartments of the same bifurcation), bif tilt remote (the angle between the previous father node of the current bifurcating father and its two daughter nodes), bif torque local (the angle between the plane of previous bifurcation and the current bifurcation), bif torque remote (the angle between current plane of bifurcation and previous plane of bifurcation), helix (Helicity of the branches of the neuronal tree. Needs to be at least 3 compartments long to compute the helicity, computes the normal form on the 3 vectors to find the 4th vector. $\frac{(\overline{A}x\overline{B})x\overline{C}}{3(|\overline{A}||\overline{B}||\overline{C}|)}$), fractal dim (the slope of linear fit of regression line obtained from the log-log plot of path distance vs euclidean distance), partition asymmetry ($\frac{|n_1-n_2|}{n_1+n_2-2}$, n_1 is the number of tips on the left and n_2 on the right)

Figura 2.5: Atributos de L-Measure (IV)

2.3. Programa Java y Base de Datos mySQL

Hemos implementado un programa java que lee las neuronas descargadas de NeuroMorpho de sus correspondientes carpetas (specie, gender, cell_type_level_2, brain_region, development y neocortex) y calcula las medidas L-Measure generando los correspondientes ficheros. Posteriormente lee de los ficheros generados los atributos y los almacena en la base de datos junto con los datos de la neurona y las etiquetas de las clases correspondientes.

Una vez que tenemos los datos disponibles en la base de datos podemos leer las clases y las instancias de las neuronas deseadas utilizando Weka y generar los ficheros *.arff* que salvaremos para posteriormente entrenar con dichos ficheros los clasificadores.

Hemos automatizado con java y la api de Weka la ejecución de las 10 iteraciones y sus

correspondientes clasificadores, además de la selección de atributos cfs y wrapper. Debemos situamos los ficheros *.arff* en la carpeta correspondiente (wekaFiles) y ejecutar el programa. Se generan los distintos clasificadores: sin selección de atributos, con selección de atributos CFS y con selección de atributos wrapper. Los atributos seleccionados y la matriz de confusión y los porcentajes de precisión se generan en la misma carpeta en dos ficheros con los nombres:

{nombre del fichero arff}_subsample_{tipo de selección de atributos}_{número de iteración}_{nombre del clasificador}

{nombre del fichero arff}_attr_subsample_{tipo de selección de atributos}_{número de iteración}_{nombre del clasificador}

Capítulo 3

Tratamiento de los Datos

3.1. Limpieza de datos erróneos

En la base de datos en general las neuronas están parcialmente completas. La mayoría incluyen el soma pero existen algunas como por ejemplo, las neuronas sensoriales en las que dada su estructura no aparece el soma. Además están las formadas únicamente por el botón sináptico que se estudia separado del resto de la neurona por ser una parte muy importante de la misma. A priori vamos a considerar todas las neuronas y excluiremos únicamente aquellas con datos erróneos.

Para evitar la pérdida de información por la naturaleza de las características estudiadas sólo vamos a tratar neuronas en tres dimensiones. La fórmula utilizada para calcular el volumen con el software L-Measure permite que una neurona en dos dimensiones pueda contener volumen a pesar de ser una suposición errónea. Por tanto filtramos las características de altura, anchura y profundidad (x, y, z) para las que el valor es 0. Borramos 452 neuronas que tienen profundidad 0. Existen dos neuronas del gusano c-elegans, una con anchura 0 y otra con profundidad 0 que vamos a conservar porque al revisar la gráfica 3D comprobamos que son correctas.

Además hemos eliminado una neurona con el tamaño del soma negativo porque esto no puede darse en la vida real. Y tres neurona que el software L-Measure no es capaz de tratar generando errores. Los números resultantes se muestran en la tabla 3.1.

Neuronas con soma	9481
Neuronas sin soma	716
Botón sináptico	682
Neuronas válidas totales	10879
Neuronas eliminadas 2D	452
Neuronas eliminadas por datos erróneos	4
Neuronas totales de NeuroMorpho	11335

Tabla 3.1: Neuronas extraídas de NeuroMorpho

3.2. Datos no balanceados

Un conjunto de datos no está balanceado cuando las distintas categorías no están representadas con una cantidad similar de datos. Por la naturaleza de los datos en NeuroMorpho encontramos categorías con más de tres mil neuronas, por ejemplo en las especies rata y ratón que son muy utilizados en los estudios de laboratorio, frente a otras especies con apenas una o dos neuronas como vemos en la figura 5.1.

Trabajar con clases no balanceadas puede producir los resultados de sobreajuste que vemos en la figura 3.1 (a) donde asignamos a neuronas de otras especies las dos clases más representativas. Esto se debe a que el algoritmo en el entrenamiento se ha ajustado demasiado a los datos pertenecientes a estos dos grandes grupos.

```
=== Confusion Matrix === All>30 wrapper best first

   a    b    c    d    e    f    g    h    i    j    k    l    m    n    o    p    q   <-- classified as
2892  430   60   13   19    6    0    3    1    0    7    6    0    0    2    0    2 |    a = Rat
 226 2610   16    0   12    9    0   10    1    0    3    4    2    0   11    6    1 |    b = Mouse
   3   22 1024    0    0    0    0    8    0    0    0    1    0    2    0    0    7 |    c = Drosophila
  20   12    0 2070   38    0    0    0    0    0    0    7    0    0    0    0    0 |    d = Human
  14   13    0    1  341    0    3    1    0    0    3    0    0    0    0    0    0 |    e = Monkey
  18   19    0    0   20   45    0    1    0    0    0    0    0    0    0    0    0 |    f = Cat
   0   25    0    0    0    0    8    1    0    0    0    0    0    0    0    0    0 |    g = Chicken
  29   36    1    0    0    0    0   34    0    0    0    0    0    0    0    0    0 |    h = Goldfish
   1    2    0    2    1    0    0    0   70    0    0    0    0    0    0    0    0 |    i = Elephant
   5   11    6    0    0    0    0   10    0    0    0    0    0    0    0    0    0 |    j = Zebrafish
   1    5    0    0    0    0    0    0    0    0  194    0    0    0    0    0    0 |    k = Sheep
   1    2    1    0    1    0    0    0    0    0    0  297    0    0    0    0    0 |    l = C_Elegans
   0    4    1    0    0    0    0    0    0    0    0    0   25    0    0    0    0 |    m = Dragonfly
   0    0    0    0    0    0    0    0    0    0    0    0    0   81    0    0    0 |    n = Agouti
  24    2    0    0   13    0    0    0    0    0    0    0    0    0   25    0    0 |    o = Salamander
   3    4    0    0    0    0    0    0    0    0    0    0    0    0    1   48    0 |    p = Blowfly
   0    0   10    0    0    0    0    1    0    0    0    0    0    2    0    0  114 |    q = Cricket
```

(a) clasificación por especies

```
Percent correct: 80.39823008849558
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         1          0.804      1         0.891      0.497    Male
                   0         0          0          0         0          0.497    Female
Weighted Avg.    0.804     0.804      0.646      0.804     0.717      0.497

=== Confusion Matrix ===

   a    b   <-- classified as
1817    0 |    a = Male
 443    0 |    b = Female
```

(b) clasificación por géneros para la especie rata

Figura 3.1: Matriz de confusión de la clasificación de neuronas con datos no balanceados

En la figura 3.1 (b) la precisión de la predicción no es apropiada porque aunque se han clasificado correctamente más del 80 % de los datos, en realidad se están asignando todos los datos a la clase macho y ninguno a la clase hembra.

Existen diversos métodos para solucionar este problema resumidos en Chawla (2005). Principalmente se dividen en dos: sobremuestreo y submuestreo. En el submuestreo se eliminan

instancias de la clase que presenta mayor número de ellas, por el contrario en el sobremuestreo se duplican o generan instancias de la clase que presenta menor número para obtener el conjunto de datos balanceado.

Después de realizar algunas pruebas con sobremuestreo los resultados estaban sesgados y obteníamos predicciones cercanas al 100 % en las clases donde las muestras habían sido duplicadas, y aunque la técnica de SMOTE propuesta por Chawla *et al.* (2002) se recomienda como la más apropiada, en nuestro caso la diferencia en la representación entre las distintas categorías supone generar cerca del 90 % de las instancias. Hemos preferido mantener los datos de neuronas que encontramos en NeuroMorpho para ajustarnos a la realidad y evitar introducir ruido que pueda corromper las muestras existentes.

Por ello hemos utilizado el borrado aleatorio de instancias (submuestreo) para balancear los datos en el que cada instancia de las clases mayoritarias tiene la misma probabilidad de ser borrada. Este filtrado puede eliminar instancias potencialmente relevantes y existen mejoras como la propuesta de Kubat *et al.* (1997) donde se borra en una dirección (one-sided selection) muestras alejadas del borde de decisión de la clase que tiene más instancias. Para evitar la sensibilidad al ruido del anterior trabajo Laurikkala (2001) se centra en la limpieza de los datos en vez de en el borrado. Utiliza el vecino más cercano de edición de Wilson para eliminar instancias de la clase con más datos. Estas mejoras no están disponibles en la versión 3.6.10 de la *api* de Weka y no las hemos utilizado.

Hemos creado grupos con las neuronas descartando las categorías que tienen menos de 30 instancias por ser grupos sin información suficiente. Al tomar la decisión de balancear las neuronas a 30 con el borrado aleatorio perdemos gran cantidad de datos. Para resolverlo hemos tomado dos caminos: en primer lugar hemos creado otros dos grupos con muestras a partir de 70 y de 800, lo que supone reducir el número de clases aumentando el número de instancias. En segundo lugar hemos ejecutado por cada grupo diez borrados aleatorios con su correspondiente clasificador y en los resultados finales se muestra la media de los diez modelos.

En los datos balanceados con el borrado aleatorio se han resuelto los problemas descritos que producen los datos sin balancear como vemos en la figura 3.2.

```
=== Confusion Matrix === All>30 subsample wrapper best first

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q   <-- classified as
 14  6  0  0  0  0  0  3  1  1  2  1  0  0  2  0  0 |  a = Rat
  2 20  1  0  0  0  2  2  0  1  0  0  1  0  1  0  0 |  b = Mouse
  0  0 29  0  0  0  0  0  0  0  0  0  0  1  0  0  0 |  c = Drosophila
  0  1  0 27  2  0  0  0  0  0  0  0  0  0  0  0  0 |  d = Human
  0  0  0  0 25  2  0  1  0  0  1  0  0  0  1  0  0 |  e = Monkey
  0  5  0  0  0 23  1  1  0  0  0  0  0  0  0  0  0 |  f = Cat
  0  0  0  0  0  0 27  1  0  1  0  0  0  0  1  0  0 |  g = Chicken
  0  0  0  0  0  0  0 19  0  9  0  0  0  2  0  0  0 |  h = Goldfish
  1  1  0  0  0  0  0  1 26  0  0  0  0  0  1  0  0 |  i = Elephant
  0  0  0  0  0  0  0  1  0 23  0  2  0  4  0  0  0 |  j = Zebrafish
  0  0  0  0  0  0  0  0  0  0 30  0  0  0  0  0  0 |  k = Sheep
  0  0  0  0  0  0  0  0  0  0  0 30  0  0  0  0  0 |  l = C_Elegans
  0  0  0  0  0  0  0  0  0  0  0  0 30  0  0  0  0 |  m = Dragonfly
  0  0  0  0  0  0  0  0  0  0  0  0  0 30  0  0  0 |  n = Agouti
  0  0  0  0  0  0  0  0  0  0  0  0  0  0 30  0  0 |  o = Salamander
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 30  0 |  p = Blowfly
  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0 29 |  q = Cricket
```

(a) clasificación por especies balanceando a 30 neuronas por clase

```
Percent correct: 90.51918735891648
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.88      0.07       0.926     0.88      0.903      0.963    Male
                 0.93      0.12       0.886     0.93      0.907      0.963    Female
Weighted Avg.    0.905     0.095      0.906     0.905     0.905      0.963
=== Confusion Matrix ===

   a   b   <-- classified as
 390  53 |   a = Male
  31 412 |   b = Female
```

(b) clasificación por géneros para la especie rata balanceado a 443 neuronas por clase

Figura 3.2: Matriz de confusión de la clasificación de neuronas con datos balanceados

3.3. Selección de atributos

Hemos utilizado las características obtenidas con el software L-Measure. Son un total de 215 e incluyen la suma, la media, el mínimo, el máximo y la desviación estándar descritas en la sección 2.2. En teoría, cuanto mayor es el número de características mejor deben ser los resultados de clasificación, pero añadir atributos irrelevantes o redundantes puede confundir a los algoritmos de aprendizaje automático. Hemos hecho selección de atributos para obtener un subconjunto de los originales y eliminar este efecto negativo.

En los algoritmos de aprendizaje automático los atributos redundantes e irrelevantes producen distintos efectos como se describe en Witten y Frank (2005): los clasificadores basados en instancias como el *IB1* son muy sensibles a los atributos irrelevantes porque trabajan con vecinos locales, tomando grupos de instancias en cada decisión; en cambio la *red bayesiana naïve* al no fragmentar el espacio de instancias ignora los atributos irrelevantes porque se asume por diseño que todos ellos son condicionalmente independientes dada la variable clase. Sin embargo es muy sensible a atributos redundantes porque si dos o mas atributos están correlados recibirán demasiado peso en la decisión final sobre a qué clase pertenecen. En el árbol

de clasificación C4.5 un atributo irrelevante se elegirá en algún momento para la ramificación del árbol, provocando un deterioro en la precisión de los resultados. A pesar de que se elige el mejor atributo en cada nodo, llegará un momento en el que teniendo pocos datos el atributo irrelevante parecerá suficientemente bueno para seleccionarlo. En la máquina de vectores soporte todos los atributos tienen el mismo peso en el cálculo de la distancia, por tanto si tenemos atributos irrelevantes se generarán errores en el cálculo del hiperplano de margen máximo.

Además al reducir las dimensiones se mejora la eficiencia y la interpretación del modelo que nos permite centrarnos en los atributos más relevantes. En Saeys *et al.* (2007) encontramos una revisión de los métodos disponibles para dicha selección. Es evidente que por la naturaleza de nuestros datos existen atributos redundantes producidos por introducir sobre una misma característica el total, la media, el máximo y el mínimo que en algunos casos tendrán relación directa e incluso llegarán a coincidir como ocurre, por ejemplo con el número de ramas que salen del soma, donde sólo se debe considerar el valor total de la suma.

Hay dos enfoques principales a la hora de seleccionar los atributos. El primero es el *método de filtrado o filter* donde se filtra antes del aprendizaje para reducir las características a aquellas más prometedoras. En cambio en el segundo método llamado *método de envoltura o wrapper* se evalúan las características con el algoritmo de clasificación que será utilizado posteriormente para el aprendizaje.

En nuestras pruebas hemos utilizado los dos enfoques:

a) El filtro multivariante proporcionado por Weka *CfsSubsetEval* propuesto por Hall (1999b), que evalúa la capacidad de cada atributo individualmente para determinar la clase y la redundancia entre los distintos atributos, seleccionando aquellos que están altamente correlados con la clase pero tienen baja intercorrelación entre si.

b) El método de envoltura *WrapperSubsetEval* propuesto por Kohavi y John (1997) que evalúa los subconjuntos de atributos utilizando un clasificador y validación cruzada para estimar su bondad en cada subconjunto. A pesar de ser menos eficiente que el anterior en tiempo de ejecución nos ha dado en general resultados más precisos.

En ambos casos hemos utilizado el método de búsqueda *BestFirst* que busca en el espacio de los subconjuntos de atributos utilizando el algoritmo voraz *Hill-Climbing* con retroceso. Hemos utilizado los valores por defecto. Por tanto, el algoritmo comienza con un conjunto vacío de atributos buscando hacia adelante y genera todas las posibles expansiones de atributos individuales. Se elige el subconjunto que maximiza el resultado de evaluación y se expande de la misma manera, añadiendo atributos individuales. Si al expandir el subconjunto no mejora, se prueba con el siguiente subconjunto mejor. Con suficiente tiempo este algoritmo evalúa todo el espacio de atributos, pero suele limitarse la búsqueda con un criterio de parada: cinco expansiones completas sin mejorar los resultados (Hall (1999a)).

Además de aplicar selección de atributos para clasificar, hemos utilizado también el conjunto completo de los atributos para poder comparar la precisión de los modelos. Pueden verse los resultados en las gráficas del anexo.

A continuación mostramos el esquema de los estudios realizados para los 4 clasificadores considerados: red bayesiana naïve, árbol de clasificación C4.5, IB1 y el clasificador máquina

de vectores soporte.

$$\text{Clasificadores por especie} \begin{cases} \text{Todas las neuronas} \begin{cases} \text{4 clases, 800 neuronas por clase} \\ \text{12 clases, 70 neuronas por clase} \\ \text{17 clases, 30 neuronas por clase} \end{cases} \\ \text{Células principales} \begin{cases} \text{3 clases, 800 neuronas por clase} \\ \text{8 clases, 70 neuronas por clase} \\ \text{13 clases, 30 neuronas por clase} \end{cases} \\ \text{Interneuronas} \begin{cases} \text{5 clases, 70 neuronas por clase} \\ \text{4 clases, 30 neuronas por clase} \end{cases} \\ \text{Terminales axónicos} \begin{cases} \text{5 clases, 70 neuronas por clase} \\ \text{6 clases, 30 neuronas por clase} \end{cases} \end{cases} \tag{3.1}$$

$$\text{Clasificadores por género} \begin{cases} \text{Todas las especies} \begin{cases} \text{Todas las neuronas} \\ \text{Células principales} \\ \text{Interneuronas} \\ \text{Terminales axónicos} \end{cases} \\ \text{Especie humana} \begin{cases} \text{Células principales} \end{cases} \\ \text{Especie rata} \begin{cases} \text{Todas las neuronas} \\ \text{Células principales} \end{cases} \end{cases} \tag{3.2}$$

$$\text{Clasificadores por edad} \begin{cases} \text{Especie rata} \begin{cases} \text{Todas las neuronas} \\ \text{Células principales} \\ \text{Interneuronas} \end{cases} \\ \text{Especie ratón} \begin{cases} \text{Todas las neuronas} \\ \text{Células principales} \\ \text{Interneuronas} \end{cases} \\ \text{Especie humana} \begin{cases} \text{Células principales} \end{cases} \\ \text{Especie drosophila} \begin{cases} \text{Todas las neuronas} \\ \text{Terminales axónicos} \end{cases} \end{cases} \tag{3.3}$$

$$\text{Clasificadores por tipo de célula} \begin{cases} \text{Todas las neuronas} \\ \text{Células principales} \\ \text{Interneuronas} \\ \text{Terminales axónicos} \end{cases} \tag{3.4}$$

$$\text{Clasificadores por región del cerebro: general} \begin{cases} \text{Todas las neuronas} \begin{cases} \text{3 clases, 800 neuronas por clase} \\ \text{15 clases, 70 neuronas por clase} \\ \text{21 clases, 30 neuronas por clase} \end{cases} \\ \text{Células principales} \begin{cases} \text{3 clases, 800 neuronas por clase} \\ \text{11 clases, 70 neuronas por clase} \\ \text{16 clases, 30 neuronas por clase} \end{cases} \\ \text{Interneuronas} \begin{cases} \text{11 clases, 70 neuronas por clase} \\ \text{5 clases, 30 neuronas por clase} \end{cases} \\ \text{Terminales axónicos} \begin{cases} \text{3 clases, 70 neuronas por clase} \\ \text{6 clases, 30 neuronas por clase} \end{cases} \end{cases} \tag{3.5}$$

$$\text{Clasificadores por región del cerebro: neocórtex} \begin{cases} \text{Todas las neuronas} \begin{cases} \text{15 clases, 70 neuronas por clase} \\ \text{19 clases, 30 neuronas por clase} \end{cases} \\ \text{Células principales} \begin{cases} \text{15 clases, 70 neuronas por clase} \\ \text{19 clases, 30 neuronas por clase} \end{cases} \\ \text{Interneuronas} \begin{cases} \text{6 clases, 30 neuronas por clase} \end{cases} \end{cases} \tag{3.6}$$

Capítulo 4

Clasificación Supervisada

Un clasificador es una función f que mapea un vector de características de entrada $\mathbf{x} = (x_1, x_2, ..., x_n) \in \chi$ y obtiene como salida la etiqueta de la clase $c \in \{1, ..., C\}$, donde χ es el espacio de características. Normalmente asumimos que $\chi = \Re^D$ o $\chi = \{0, 1\}^D$. En nuestro caso el vector de características contiene tanto valores discretos, por ejemplo el número de ramificaciones que salen del soma, como continuos, por ejemplo el volumen de los compartimentos. Además se asume que las etiquetas de clase son mutuamente excluyentes. El objetivo es aprender la función f de un conjunto de datos de entrenamiento.

Hemos elegido de entre los algoritmos de aprendizaje automático un clasificador probabilístico: red bayesiana naïve y tres clasificadores no probabilísticos: un algoritmo basado en instancias (IB1), un árbol de decisión (C4.5) y el clasificador máquina de vectores soporte que combina los métodos basados en instancias con los modelos lineales. Esta variedad nos permitirá hacer una comparación de la precisión de los resultados y su eficiencia computacional tanto en memoria como en tiempo entre los distintos algoritmos.

4.1. Clasificadores

4.1.1. Clasificador bayesiano naïve

Los clasificadores bayesianos (Bielza y Larrañaga (2014)) asignan la clase más probable dado un ejemplo descrito por su vector de características. El clasificador bayesiano naïve es un clasificador probabilístico que asume la independencia condicional entre los atributos dada la clase:

$$p(\mathbf{X}{=}\mathbf{x}|C = c) = \prod_{i=1}^{n} p(X_i = x_i|C = c) \tag{4.1}$$

donde $\mathbf{x}$ es el vector de atributos y c es la clase. Aunque la suposición de independencia es una asunción un poco pobre, en la práctica se obtienen muy buenos resultados. En la figura 4.1 vemos la estructura de la red.

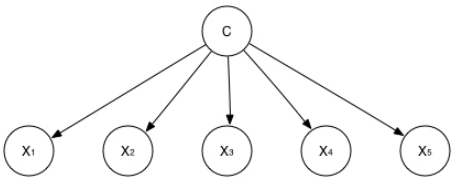


Figura 4.1: Estructura de red bayesiana naïve

La clasificación c^* se realiza mediante

$$c^* = arg\ max\ p(c|x) = arg\ max\ p(c) \prod_i p(x_i|c) \tag{4.2}$$

Entre los algoritmos de Weka encontramos NaïveBayesSimple propuesto por Duda *et al.* (1999), que utiliza la distribución normal para modelar los atributos continuos, frente a NaïveBayes propuesto por John y Langley (1995) que permite utilizar un estimador kernel de densidad no paramétrico para discretizar. Vemos las diferencias en la figura 4.2. En la estimación kernel con kernel Gausiano la densidad estimada se promedia utilizando un conjunto de kernels:

$$p(\mathbf{X}{=}\mathbf{x}|C = c) = \frac{1}{n} \sum_i g(x, \mu_i, \sigma_c) \tag{4.3}$$

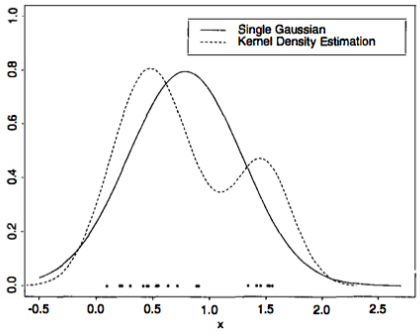


Figura 4.2: Gráfica del efecto de usar una Gausiana frente al método kernel para estimar la densidad de una variable continua (John y Langley (1995))

Además el algoritmo NaïveBayes realiza comprobaciones de *underflow* en el cálculo de las probabilidades de pertenencia a cada clase para la instancia dada. Hemos utilizado este algoritmo por dicha propiedad evitando los errores (*Can't normalize array. Sum is zero*) que obteníamos con el otro algoritmo (NaïveBayesSimple). Pero hemos mantenido la distribución Gausiana para modelar los atributos continuos.

4.1.2. Árbol de clasificación C4.5

Un árbol de clasificación es una estructura con forma de árbol donde cada nodo de decisión evalúa un atributo y cada hoja representa el valor de la clase. Para cada posible valor evaluado

tenemos un nodo hijo. Asignar a una instancia un valor de clase depende de los valores de los atributos del caso tratado, que generará un camino desde el nodo raíz hasta una hoja donde la clase especificada es el resultado predicho por el árbol.

El algoritmo C4.5 descrito en Quinlan (1993) y llamado J48 en Weka crea su estructura con la estrategia divide y vencerás donde cada nodo está asociado con un conjunto de casos. Extiende el algoritmo original ID3 propuesto también por Quinlan (1986) con las siguientes mejoras o extensiones: maneja atributos discretos y continuos, los atributos con valores perdidos los ignora y permite una poda para eliminar las partes que no aportan valor una vez que ha sido generado el árbol. Utiliza la cantidad de información mutua entre cada variable predictora X_i y la variable clase C:

$$I(C, X_i) = H(C) - H(C|X_i) \tag{4.4}$$

Para seleccionar el siguiente nodo en la construcción del árbol se escoge el atributo X_i que maximiza el criterio:

$$I(C, X_i)/H(X_i) \tag{4.5}$$

$$H(X_i) = -\sum p(x_i) log_2 p(x_i) \tag{4.6}$$

Los árboles de clasificación trabajan con atributos nominales de forma natural, una vez que se ha utilizado en una rama el atributo nominal se habrá utilizado toda la información que nos ofrece y se prueba únicamente una vez. En cambio en las divisiones con atributos numéricos se continúa produciendo información pudiendo probarlos varias veces, lo que produce árboles desordenados y complicados de entender porque las pruebas con atributos numéricos están dispersas a lo largo del camino. Una alternativa que mejora la legibilidad del árbol es permitir más de dos caminos desde un mismo nodo. Otra forma más sencilla pero menos potente es discretizar los atributos.

Los árboles completamente expandidos suelen contener información irrelevante y se simplifican con una poda una vez construido el árbol. La poda que realiza el algoritmo C4.5 a veces no es suficiente y el árbol contiene ramas innecesarias, no obstante es rápida y por ello muy popular.

4.1.3. Clasificador IB1

En el aprendizaje basado en instancias se almacenan las instancias de entrenamiento y se utiliza una función de similitud que determina que instancia del conjunto de entrenamiento es más parecida a la instancia de test de la que queremos conocer la clase. Una vez que se ha localizado el vecino mas cercano del conjunto de entrenamiento se asigna el valor de la clase a dicha instancia.

Hemos utilizado el algoritmo con los parámetros por defecto que proporciona Weka descrito en Aha *et al.* (1991) y resumido en el algoritmo 1.

Algoritmo 1: Algoritmo IB1 (CD = Descripción de concepto)

$CD \longleftarrow \emptyset$
for *each* $\boldsymbol{x} \epsilon$ *Conjunto Entrenamiento* **do**
  for *each* $\boldsymbol{y} \epsilon\ CD$ **do**
    $Sim[y] \longleftarrow Similarity(x, y)$
  if $class(x) = class(y_{max})$ **then** $classification \longleftarrow correct$
  else $classification \longleftarrow incorrect$;
  $CD \longleftarrow CD \cup \{x\}$

La descripción de concepto contiene instancias con el valor de la clase asignada. La función de similitud es la distancia euclídea normalizada y junto con la función de clasificación determinan cómo el conjunto de instancias almacenadas en la descripción de concepto se utilizan para predecir el valor de la clase.

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \tag{4.7}$$

Las instancias se describen utilizando n atributos, donde x_i es un atributo de la instancia de entrenamiento considerada y y_i es el atributo de la instancia de test. Se define $f(x_i, y_i) = (x_i - y_i)^2$ para atributos numéricos y $f(x_i, y_i) = \begin{cases} 1 & \text{si } x_i \neq y_i \\ 0 & \text{si } x_i = y_i \end{cases}$ para atributos categóricos.

IB1 se diferencia del algoritmo *nearest neighbour* o el vecino más cercano en que reduce el espacio de almacenamiento, normaliza los rangos de los atributos escalándolos de acuerdo al valor máximo y mínimo representado por el atributo, procesa las instancias incrementalmente aprendiendo tramos lineales y tiene una política de tolerancia para los atributos perdidos, donde se asume que el valor perdido tiene la máxima distancia con el que existe. Pero si faltan ambos, el atributo de la instancia de entrenamiento y el de la instancia de test se asigna el valor 1. El algoritmo considera que instancias parecidas tienen clases similares y que todos los atributos tienen la misma relevancia en la decisión de clasificación siempre que no se tiene conocimiento previo.

Los problemas de los algoritmos basados en instancias son la sensibilidad al ruido y a los atributos irrelevantes. Además son caros computacionalmente, sensibles a la función de similitud elegida y no presentan una forma natural de trabajar con valores nominales o atributos perdidos. Tampoco proporcionan información acerca de la estructura de los datos. Otro problema que presentan es que las instancias con ruido corrompen fácilmente el clasificador.

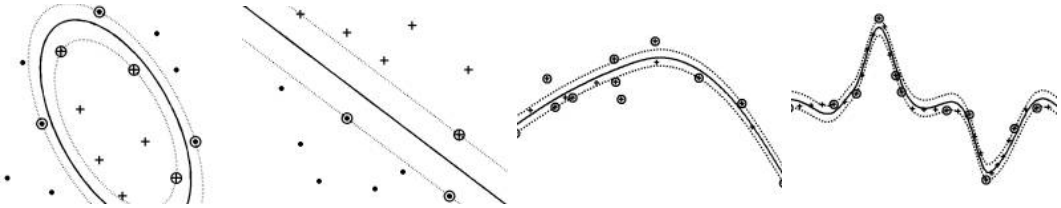
A pesar de ser un método simple y efectivo, generalmente es lento. Para encontrar el miembro del conjunto de entrenamiento más cercano a una instancia de test desconocida, se debe calcular la similitud con cada miembro del conjunto de entrenamiento y elegir el mejor resultado. Es decir, el tiempo en realizar una predicción es proporcional al número de instancias de entrenamiento.

4.1.4. Clasificador máquina de vectores soporte

El clasificador máquina de vectores soporte propuesto por Cortes y Vapnik (1995) es una mezcla entre los modelos lineales y los modelos basados en instancias. Utiliza modelos lineales

para implementar límites no lineales de la clase, transformando el espacio de instancias en uno nuevo. Con el mapeo no lineal una línea en el nuevo espacio puede representar un límite de decisión no lineal en el espacio original.

Selecciona un conjunto pequeño de instancias en los límites críticos llamados vectores de soporte de cada clase y construye una función lineal discriminante que separa las clases lo mejor posible con un hiperplano de margen máximo, definido unívocamente por el conjunto de vectores de soporte que son únicos. El hiperplano de margen máximo es el aquel que ofrece la mayor separación posible entre clases.



(a) (b) (c) (d)

Figura 4.3: Ejemplos de distintos hiperplanos de margen máximo (imágenes de *http://www.support-vector-machines.org*)

El hiperplano para dos atributos puede definirse como $x = w_0 + w_1 a_1 + w_2 a_2$ donde a_i son los valores de los atributos y se deben aprender los pesos w_i. Pero podemos escribir la ecuación en términos de los vectores de soporte:

$$f(x) = b + \sum_{i \ es \ vector \ soporte} \alpha_i \phi(x) \cdot \phi(x_i) \tag{4.8}$$

donde y_i es el valor de la clase de las instancias de entrenamiento, b y α_i son parámetros numéricos que deben ser determinados por el algoritmo y x es el vector de las instancias de test. Los vectores x_i son los vectores de soporte. Matemáticamente cualquier función $K(x, y)$ es una función kernel si puede escribirse como $K(x, y) = \phi(x) \cdot \phi(y)$ donde ϕ es una función que mapea una instancia en un espacio de características (potencialmente multidimensional). Es decir la función kernel representa el producto vectorial en el espacio creado.

Hemos utilizado la función de base radial para mapeo de límites no lineales que proporciona Weka como función por defecto:

$$\exp(-\gamma ||u - v||^2) \tag{4.9}$$

Excepto en casos triviales no es posible determinar a priori si el espacio es linealmente separable. Se asume que los datos de entrenamiento son linealmente separables en el espacio de instancias o en el nuevo espacio creado con el mapeo no lineal. Pero el clasificador máquina de vectores soporte puede generalizarse al caso no separable utilizando un límite superior en los coeficientes α_i, aunque el parámetro debe ser elegido a mano.

Comparado con otros métodos como por ejemplo los árboles de decisión, este clasificador es lento trabajando con límites no lineales, debido al número de coeficientes que se introducen en la transformación. Pero produce clasificadores muy precisos gracias a la complejidad que puede obtener en los límites de decisión. Puede que se presente sobre-ajuste si el número de

coeficientes es grande en relación al número de instancias de entrenamiento, de manera que el modelo resultante será 'no-lineal' en exceso, sobreajustando los datos de entrenamiento porque hay demasiados parámetros en el modelo. Esto crea inestabilidad porque si cambiamos uno o dos vectores de instancias provocarán grandes cambios en los bordes de decisión, pero es poco probable que ocurra porque los vectores de soporte son representantes globales de todo el conjunto de entrenamiento, y en general hay pocos.

4.2. Medidas de precisión

Como se describe en Fawcett (2006), dado un clasificador y una instancia, si la instancia es positiva y se clasifica como positiva es un verdadero positivo, si se clasifica como negativa es un falso negativo. Si la instancia es negativa y se clasifica como positiva será un falso positivo y verdadero negativo si es clasificada negativa. Dado un clasificador podemos construir su matriz de confusión, la cual contiene el número de instancias de test clasificadas correcta e incorrectamente, como vemos en la figura 4.4.

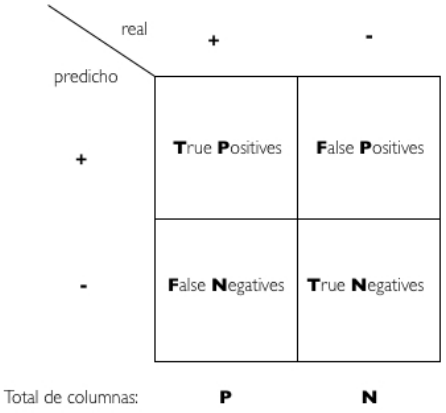


Figura 4.4: Matriz de confusión de un clasificador binario

A partir de la matriz obtenemos una serie de métricas para evaluar el rendimiento del clasificador descritas en las ecuaciones siguientes:

$$FP\ Rate = \frac{FP}{N} \tag{4.10}$$

$$TP\ Rate = \frac{TP}{P} \tag{4.11}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.12}$$

$$Recall = \frac{TP}{P} \tag{4.13}$$

$$F - Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4.14}$$

En predicción multiclase el resultado se muestra como una matriz de confusión con una fila y una columna para cada clase. Cada elemento de la matriz muestra el número de ejemplos de test para el cual la clase real es la fila y la predicha la columna. Un buen resultado se corresponde con números altos en la diagonal principal y pequeños, idealmente cero, fuera de la diagonal.

La curva ROC es un gráfico bidimensional que dibuja en el eje Y el valor de $TP\ Rate$ y en el eje X el valor de $FP\ Rate$. Muestra la relación entre verdaderos positivos y falsos positivos. El área bajo la curva nos permite comparar clasificadores: es un número comprendido entre 0 y 1 y cuanto más se acerque a 1 mejor será la precisión del modelo estudiado.

Una forma honesta de medir la tasa de error en un esquema de aprendizaje dado un conjunto particular de datos, es utilizar la validación cruzada 10 veces (*10 fold cross-validation*). Los datos se dividen aleatoriamente en 10 partes donde cada clase está representa aproximadamente en la misma proporción en el conjunto completo de datos. Una parte de las 10 se utiliza para testear el clasificador calculando la tasa de error, mientras que las 9 restantes se utilizan para entrenar el clasificador. El procedimiento de aprendizaje se ejecuta 10 veces en los diferentes conjuntos y las 10 estimaciones de error realizadas se promedian obteniendo la estimación total.

Capítulo 5

Resultados Obtenidos

Hemos creado grupos con todas las neuronas juntas y separando los tipos (interneuronas, terminales axónicos y células principales) que aparecen en la *web* de NeuroMorpho; pero no hemos obtenido diferencias sustanciales al hacer esta separación. La pequeña mejora observada en la precisión podría explicarse porque, en general, se reduce el número de clases. Ésta reducción es debida a que al separar, por ejemplo, la especie humana sólo contiene células principales desapareciendo en los otros dos grupos (interneuronas y terminales axónicos), esto ocurre con todas las clases que están representadas únicamente por uno de los tipos. En las siguientes secciones estudiamos en profundidad los resultados que hemos considerado más interesantes.

Se ha utilizado 10 *fold cross-validation* para evaluar los resultados. Además, como se ha comentado anteriormente, hemos obtenido por cada clasificador 10 modelos seleccionando los conjuntos de instancias aleatoriamente y aplicando a cada uno de ellos *10 fold cross-validation*. Estudiaremos aquellos modelos que han dado mejores resultados en porcentajes de acierto o aquellos en los que su reducido número de atributos justifica una disminución de aciertos inferior al 5 %. En las gráficas del anexo se muestra la media de los porcentajes obtenidos en las 10 iteraciones. Existen otras formas de seleccionar las clases, como por ejemplo utilizar los 10 modelos asignando la clase final por votación como propone Breiman (1996).

Como vimos en la sección 3.3 de datos no balanceados nuestros datos presentan mucha diferencia en la cantidad de neuronas de cada clase. Para no tener que elegir entre cantidad de instancias o número de clases hemos creado grupos donde se mantenga el número de instancias o el número de clases, reduciendo en cada caso lo que corresponda.

5.1. Clasificación de la especie

Existen investigadores que han estudiado las diferencias en la morfología de las neuronas entre las especies, especialmente entre los mamíferos por sus similitudes en otros aspectos. Purves y Lichtman (1985) han investigado las neuronas homólogas del ganglio superior cervical en mamíferos de distintos tamaños: hamsters, ratones, ratas, cerdos de Guinea y conejos. Para el estudio utilizaron varias variables que consideraron como medidas de complejidad de la neurona: número de dendritas primarias que salen del soma, longitud total de las dendritas y los procesos sinápticos excitatorios, concluyendo que la longitud y la complejidad dendrítica están relacionadas con el tamaño del animal estudiado, siendo mayor a mayor tamaño del

animal adulto. En publicaciones más recientes han llegado a conclusiones parecidas. Jacobs *et al.* (2014) han realizado un estudio profundo de los distintos tipos de neuronas del córtex cerebral en las especies: elefante africano, manatee, tigre siberiano, pantera nebulosa, ballena jorobada, jirafa, chimpancé común y humanos. Las medidas dendríticas tendían a ser mayores conforme mayor era el volumen cerebral. Entre los tipos de neuronas estudiadas (estrelladas, en cesta, Lugaro, Golgi y neuronas granuladas), en particular las neuronas Lugaro del elefante eran desproporcionalmente más largas que en otras especies, determinando que las medidas dendríticas y el tamaño del soma entre especies son diferentes.

En la figura 5.1 se muestran el número de neuronas válidas descargadas de NeuroMorpho por especie (horizontal) y tipo de neurona (vertical). En las figuras B.1, B.2, B.3 y B.4 del anexo, vemos la media de los porcentajes de acierto para los distintos clasificadores y grupos considerados. Hemos creado grupos de clasificación basándonos en la cantidad de instancias disponibles: especies con más de 30 neuronas, especies con más de 70 y especies con más de 800. De este modo evitamos perder información en aquellas especies en las que el número de instancias es considerablemente mayor.

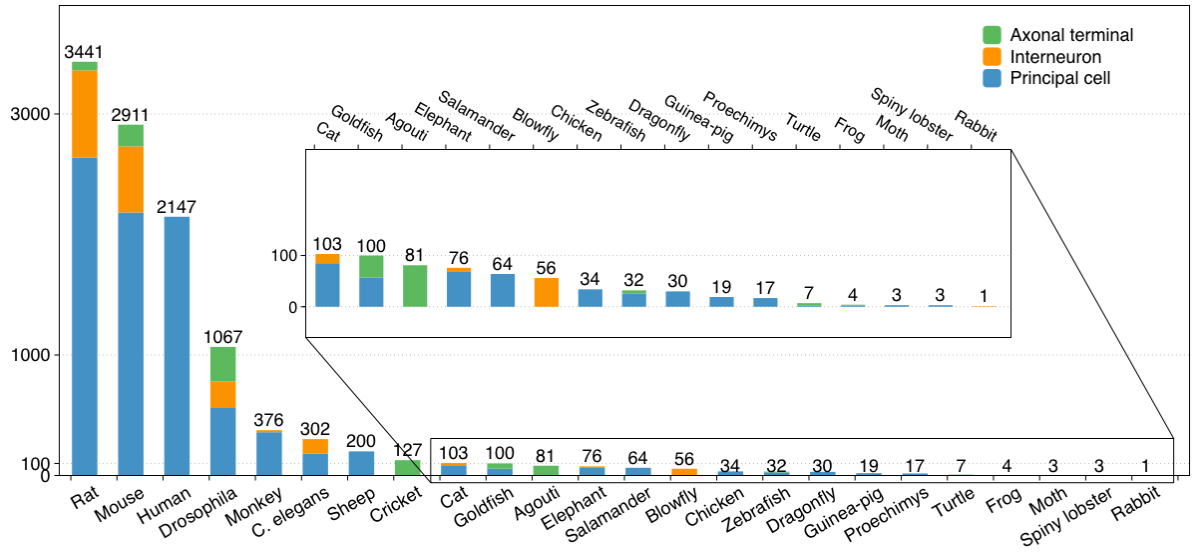


Figura 5.1: Distribución de las neuronas por especie y tipo de neurona

A continuación detallamos los grupos por tipo de neurona:

- **Grupo con todos los tipos de neuronas**: Al aplicar los modelos obtenidos para mínimo 30 instancias y 17 clases (de *rat* a *dragonfly*) a todas las neuronas disponibles obtenemos porcentajes de acierto cercanos al 60 % aunque los clasificadores bayesiano naïve e IB1 a priori parecían prometedores por sus resultados de precisión alrededor del 85 % (5.1). Parece que con 30 instancias por clase los clasificadores están sobreajustados a los datos de entrenamiento y son necesarias más instancias para mejorar los resultados. Por eso evaluaremos los resultados del clasificador de 70 instancias por especie (de *rat* a *elephant*). En la figura 5.2 (a) tenemos los resultados del modelo IB1 obtenido y en la figura 5.2 (b) la matriz resultante de aplicar el modelo al resto de datos. Vemos

en la figura resaltadas las confusiones entre las especies rata con ratón y humanos con monos lo que parece razonable, con una precisión del 90,35 %.

17 clases 30 instancias por clase (de rat a dragonfly)	**Modelo**	Correctly Classified Instances: 437	85.68 %
		Incorrectly Classified Instances: 73	14,31 %
	Total de datos	Correctly Classified Instances: 6885	61.76 %
		Incorrectly Classified Instances: 4262	38.23 %
12 clases 70 instancias por clase (de rat a elephant)	**Modelo**	Correctly Classified Instances: 824	90,35 %
		Incorrectly Classified Instances: 88	9.64 %
	Total de datos	Correctly Classified Instances: 8154	74.59 %
		Incorrectly Classified Instances: 2777	25.40 %
4 clases 800 instancias por clase (de rat a drosophila)	**Modelo**	Correctly Classified Instances: 4000	93.72 %
		Incorrectly Classified Instances: 268	6,27 %
	Total de datos	Correctly Classified Instances: 9016	94.25 %
		Incorrectly Classified Instances : 550	5.74 %

Tabla 5.1: Tabla de resultados del clasificador IB1 variando el número de instancias y clases

En la curva ROC de la figura 5.3(a) mostramos los resultados para la especie humana vs el resto de especies para el clasificador IB1 variando el número de instancias. Como puede observarse por las curvas, a medida que aumentamos el número de neuronas y disminuimos el número de clases mejoran los resultados del clasificador. En la figura 5.3(b) hemos comparado los cuatro clasificadores: bayesiano naïve, C4.5, IB1 y SVM para el grupo de 70 instancias y 12 clases. Aunque por la curva puede parecer mejor el clasificador bayesiano naïve, al aplicar todas las neuronas disponibles funciona mejor el modelo IB1 (5.1).

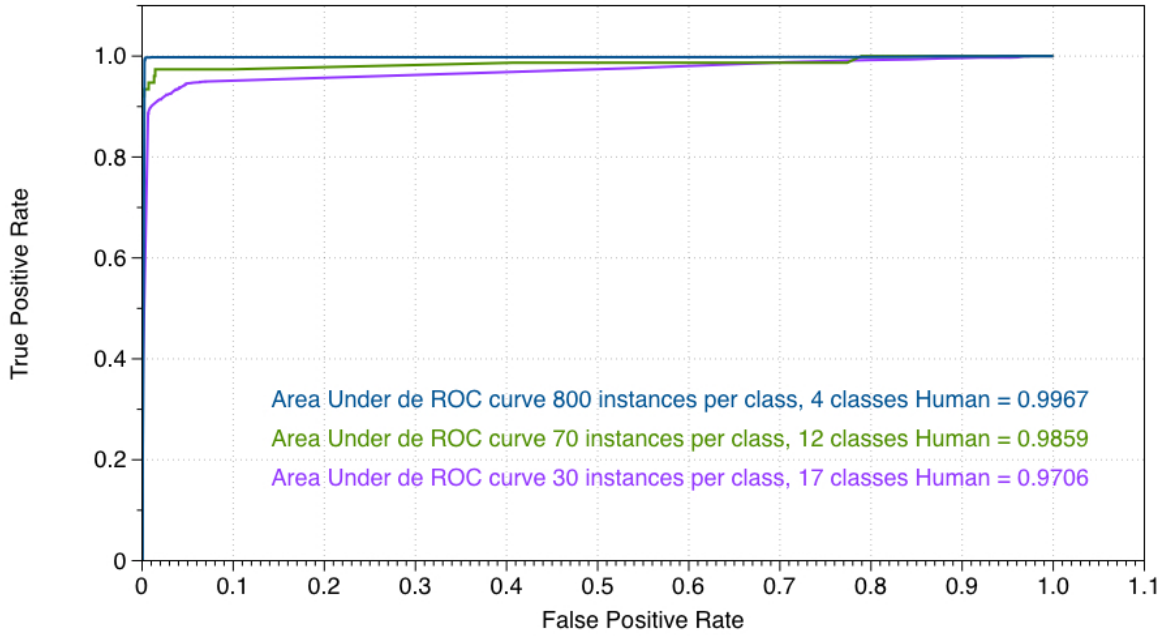
```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 55  3  0  2  2  6  2  1  3  2  0  0 |  a = Rat
  8 54  2  2  2  4  2  1  1  0  0  0 |  b = Mouse
  0  2 68  0  0  0  4  0  0  0  0  2 |  c = Drosophila
  0  1  0 71  4  0  0  0  0  0  0  0 |  d = Human
  3  0  0  3 65  2  0  0  3  0  0  0 |  e = Monkey
  3  2  0  1  1 65  2  1  1  0  0  0 |  f = Cat
  0  1  1  0  0  1 72  0  0  0  1  0 |  g = Goldfish
  2  2  0  0  0  0  0 72  0  0  0  0 |  h = Elephant
  0  0  0  0  0  0  0  0 76  0  0  0 |  i = Sheep
  0  0  0  0  0  0  0  0  0 76  0  0 |  j = C_Elegans
  0  0  0  0  0  0  0  0  0  0 76  0 |  k = Agouti
  0  0  0  0  0  0  1  0  0  0  1 74 |  l = Cricket

Correctly Classified Instances         824               90.3509 %
Incorrectly Classified Instances        88                9.6491 %
```

(a) Matriz de confusión del modelo

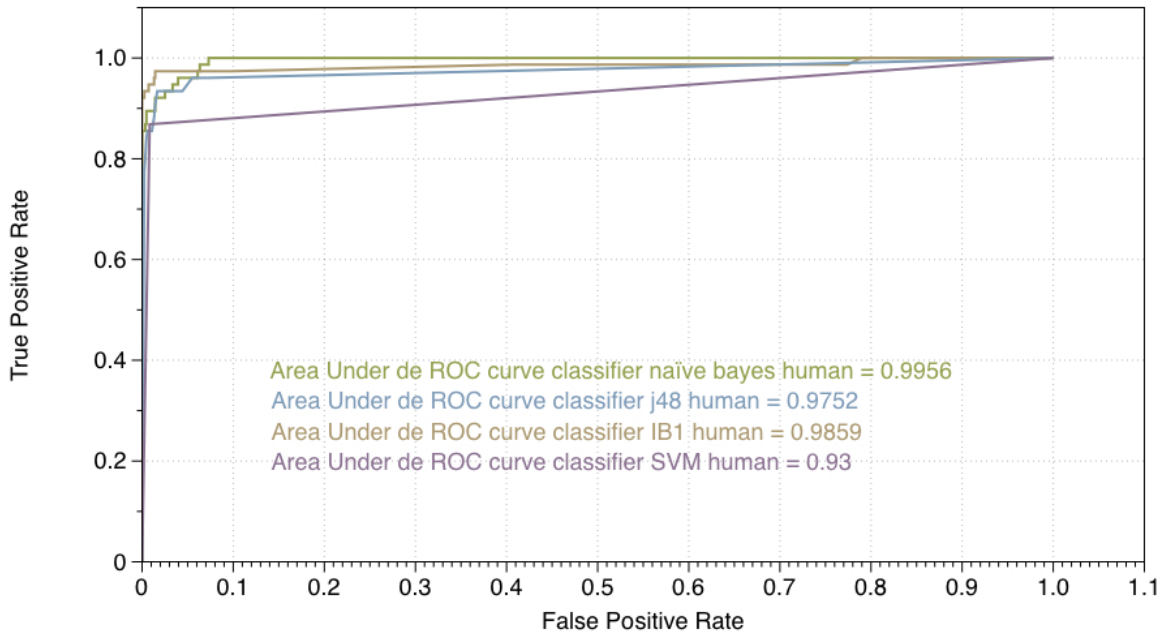
```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l   <-- classified as
 2313  358   16  118   56  196  257   24   37   64    2    0 |    a = Rat
  543 1590   35   81   51  235  182   21   92   79    2    0 |    b = Mouse
    1    1  939    0    0    0   68    0    0    3    2   53 |    c = Drosophila
   17   21    0 2007   77    1    1   15    1    7    0    0 |    d = Human
    7    5    1   11  333    9    1    0    9    0    0    0 |    e = Monkey
    2    2    0    1    0   98    0    0    0    0    0    0 |    f = Cat
    0    0    0    0    0    0  100    0    0    0    0    0 |    g = Goldfish
    0    0    0    0    0    0    0   76    0    0    0    0 |    h = Elephant
    3    2    0    0    0    0    0    1  194    0    0    0 |    i = Sheep
    3    0    0    0    0    0    0    0    0  299    0    0 |    j = C_Elegans
    0    0    0    0    0    0    0    0    0    0   81    0 |    k = Agouti
    0    0    1    0    0    0    2    0    0    0    0  124 |    l = Cricket

Correctly Classified Instances        8154               74.5952 %
Incorrectly Classified Instances      2777               25.4048 %
```

(b) Matriz de confusión del total de datos

Figura 5.2: Matrices de confusión del clasificador IB1 por especies para 12 clases



(a) Comparación de la curva ROC para la especies humana vs no humana del clasificador IB1 con 30 instancias, 70 instancias y 800 instancias



(b) Comparación de la curva ROC para la especies humana vs no humana de los distintos clasificadores: bayesiano naïve, árbol de clasificación C4.5, IB1 y SVM

Figura 5.3: Curva ROC

En la tabla 5.2 resumimos los atributos resultantes del modelo IB1 y selección de atributos wrapper. Confirmamos las medidas utilizadas por Purves y Lichtman (1985) y Jacobs *et al.* (2014), al obtener entre los atributos el número de dendritas que salen del soma, la longitud o profundidad total de la neurona y longitud, área y volumen de

los compartimentos. Además de las anteriores hemos obtenido medidas como el tipo de dendritas de la neurona, los ángulos de las bifurcaciones y relaciones entre los diámetros de los compartimentos.

Número de Atributo	**Nombre de Atributo**	**Medida**
2	n stem	total
8	depth total	total
52, 95, 138	type	avg, max, min
80, 123	bif torque local	avg, max
108	taper 2	max
110	contraction	max
115	rall power	max
125, 168	last parent diam	max, min
141	length	min
143	section area	min
144	volume	min
170	hillman threshold	min
205	bif ampl local	sd
215	fractal dim	sd

Tabla 5.2: Atributos obtenidos con selección wrapper para el modelo IB1 de clasificación por especies para 12 clases

- **Terminal axónico**: En contraste con los resultados para los clasificadores entrenados con todos los tipos de neuronas donde se han encontrado dificultades para diferenciar entre si las especies de rata y ratón, al examinar el terminal axónico por separado se han obtenido muy buenos resultados. Pero al profundizar en los datos de las distintas especies vemos que los terminales pertenecen a regiones diferentes: en las ratas pertenecen a la *médula* o el *hipocampo* mientras en los ratones son del *sistema nervioso periférico*. En la mosca drosophila encontramos el *bulbo olfativo* y el *protocerebro*, en el pez *goldfish* pertenecen al *nervio óptico*, en el *agouti* al *neocortex* y en el grillo al *sistema sensorial cercal* (Murphey y Chiba (1990)). El mejor modelo se ha obtenido en una de las iteraciones del clasificador bayesiano naïve, pero no podemos asegurar si se están diferenciando especies, regiones cerebrales o ambas. No obstante en la figura 5.4 vemos los resultados de las matrices de confusión del modelo y de aplicar el modelo a todos los terminales axónicos disponibles para las especies tratadas.

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 42  0  0  0  0  0 |  a = Rat
  0 40  1  1  0  0 |  b = Drosophila
  0  0 42  0  0  0 |  c = Goldfish
  0  0  0 42  0  0 |  d = Mouse
  0  0  0  0 42  0 |  e = Agouti
  0  0  0  0  0 42 |  f = Cricket

Correctly Classified Instances         250               99.2063 %
Incorrectly Classified Instances         2                0.7937 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

  a   b   c   d   e   f   <-- classified as
 69   0   1   0   0   0 |  a = Rat
 16 213  11   2   0  14 |  b = Drosophila
  0   0  42   0   0   0 |  c = Goldfish
  0   0   0  95   0   0 |  d = Mouse
  0   0   0   0  79   0 |  e = Agouti
  1   3   2   1   0 120 |  f = Cricket

Correctly Classified Instances         618               92.3767 %
Incorrectly Classified Instances        51                7.6233 %
```

(b) Matriz de confusión del total de datos

Figura 5.4: Matrices de confusión del clasificador bayesiano naïve por especies para los terminales axónicos

- **Células principales e interneuronas**: como podemos ver en la figura 5.5 mejora la precisión en todos los modelos en el grupo de las interneuronas, a diferencia de las células principales donde apenas varían los porcentajes. Pero debemos tener en cuenta que en éste grupo tenemos sólo 4 clases (ratón, rata, c. elegans y drosophila) muy diferentes entre sí, excepto ratón y rata donde se presentan las mismas confusiones que en el caso general.

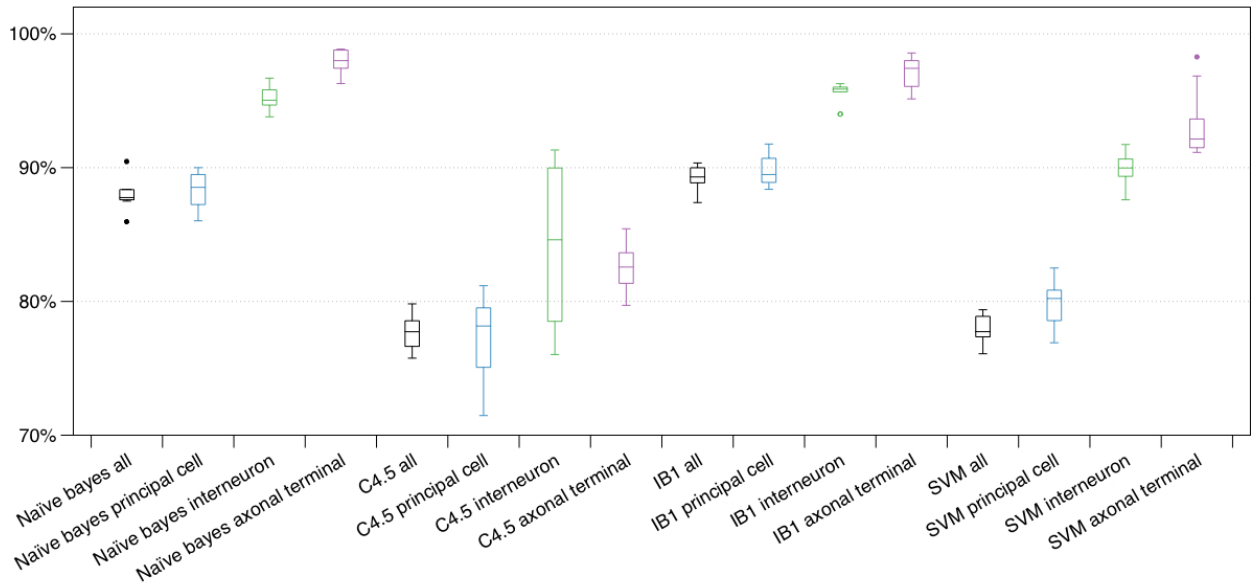


Figura 5.5: Varianza de cada clasificador para los 10 modelos con selección wrapper de clasificación de la especie para 70 instancias por clase

5.2. Clasificación del género

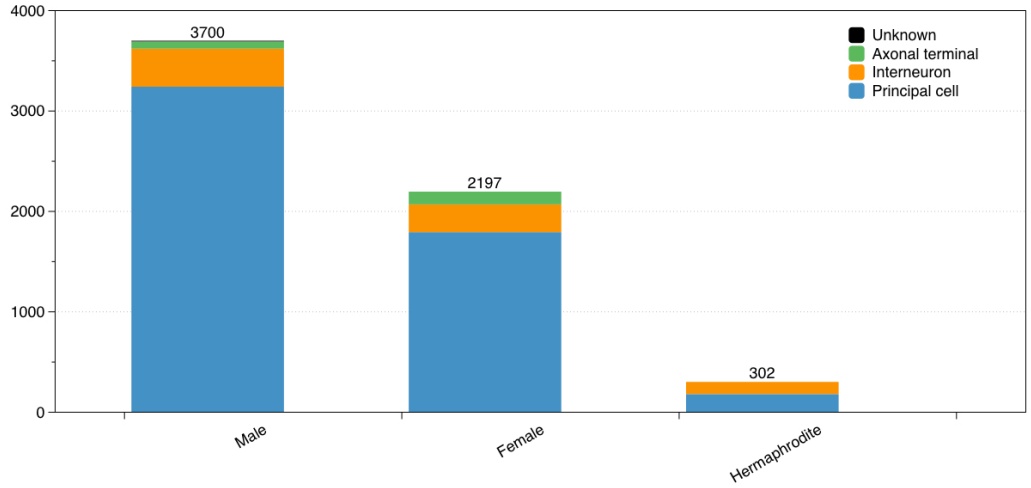
Las diferencias entre géneros o dimorfismo sexual en el cerebro se lleva estudiando desde 1980. Tobet y Fox (1992) resumen las distintas investigaciones existentes y las diferencias

encontradas en humanos a nivel de morfología cerebral, cantidad de neuronas y procesos sinápticos entre ambos sexos.

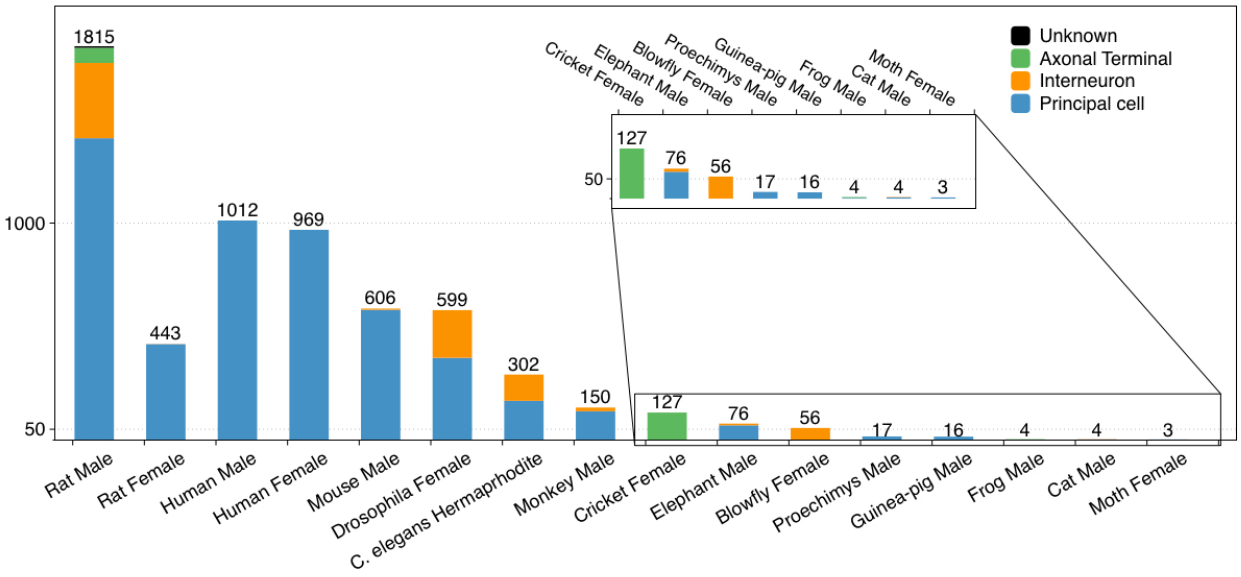
Existen también estudios para no humanos que señalan las diferencias entre géneros a nivel de morfología neuronal. Entre las investigaciones realizadas en ratas, Juraska *et al.* (1989) han publicado que en las neuronas piramidales del hipocampo, el árbol dendrítico apical cercano al soma en las hembras mostraba más material dendrítico que los machos, y viceversa según nos alejamos del soma. Markham y Juraska (2002) y Markham *et al.* (2005) afirman que en las ratas, el macho joven-adulto presenta mayor densidad de espinas y arborización en las neuronas del córtex anterior. Además, con la edad en los machos se acentúa la pérdida de densidad del árbol dendrítico resultando en una disminución del dimorfismo. Griffin y Flanagan-Cato (2009) encuentran mayor número de dendritas y longitud de las mismas en los machos en el núcleo ventromedial del hipotálamo, confirmando los resultados de Markham y Juraska (2002) y Markham *et al.* (2005). En la misma región cerebral, Dugger *et al.* (2007) observan que las ratas macho poseen un soma más grande. Los mismos resultados que han obtenido Goldstein *et al.* (1990) al realizar sus estudios en las motoneuronas del núcleo espinal bulbocavernoso conocido por ser dimórfico. Utilizaron distintas hormonas y observaron que el tamaño del soma era aparentemente menor en hembras, a pesar de intentar masculinizarlo tratando al animal con testosterona, y la longitud de las dendritas más larga en animales castrados. En cambio en la mosca drosophila, Possidente y Murphey (1989) han observado que el sexo determina la forma del axón en las neuronas sensoriales donde los axones masculinos cruzan la línea media que nunca cruzan los femeninos, influyendo el género en el tamaño del mismo. En los monos macacos Ayoub *et al.* (1983) al analizar las neuronas del área preóptica de fascicularis han comprobado que las diferencias entre sexos son exclusivas de la estructura de la neurona, es decir, no muestran diferencias en cantidad de neuronas o estructura de la zona, los machos poseen más bifurcaciones y más frecuencia de espinas. En cambio Konishi y Akutagawa (1985) encuentran más neuronas y con diámetro más largo en el núcleo del cerebro anterior en los machos del pinzón cebra.

Parece que dependiendo de la zona cerebral y la especie, el dimorfismo sexual varía, aunque en general podríamos decir que los machos presentan mayor número de espinas y mayor arborización dendrítica. Sabemos que existen zonas del cerebro que se diferencian en la cantidad de neuronas, los procesos sinápticos y la morfología de la neurona mientras otras zonas son iguales entre los distintos géneros.

Hemos clasificado las neuronas por género utilizando todas las especies juntas (figura 5.6 (a)), pero como vemos en la figura 5.6 (b), en la distribución por especie y género obtenida de NeuroMorpho, a excepción de ratas y humanos, las especies están representadas por sólo uno de los sexos. Para asegurar que realmente diferenciamos entre sexos hemos tratado por separado los grupos de especies que disponen de neuronas para ambos géneros. Los resultados obtenidos para los distintos clasificadores están en las figuras B.5, B.6, B.7 y B.8 del anexo.



(a)



(b)

Figura 5.6: Neuronas de NeuroMorpho por género

A continuación profundizamos en los mejores modelos obtenidos y los atributos utilizados en los grupos siguientes:

- **Especie Humana**: Los estudios basados en las neuronas humanas no son fáciles de realizar porque no pueden ser manipulados en laboratorios como lo son los animales. Tenemos disponibles 1981 neuronas del neocórtex, todas ellas piramidales. Gracias a la homogeneidad de las neuronas hemos tenido resultados similares en las 10 iteraciones de cada modelo.

  Si profundizamos en los resultados observamos gran diferencia en la precisión entre el clasificador máquina de vectores soporte y el resto de clasificadores (figura 5.8). Con el

primero se han obtenido porcentajes de acierto superiores al 99 % utilizando 2 atributos seleccionados con CFS, mientras que en los demás clasificadores no superan el 70 %. El máximo alcanzado con selección de atributos wrapper obtenido por la red bayesiana naïve cercano a 77 % utiliza 33 atributos. En la tabla 5.3 mostramos los atributos del mejor modelo obtenido con el clasificador máquina de vectores soporte con selección de atributos CFS: la media del valor del ángulo entre los planos de bifurcación (*bif torque remote*) y el mínimo de área de la sección de los compartimentos (*section area*) ambos valores directamente relacionados con la cantidad de material dendrítico coincidiendo con la literatura revisada en otras especies.

Número de Atributo	**Nombre de Atributo**	**Medida**
81	bif torque remote	avg
143	section area	min

Tabla 5.3: Atributos obtenidos con selección de atributos CFS para el modelo SVM de clasificación del género para la especie humana

En la figura 5.7 tenemos la matriz de confusión del modelo y la matriz de aplicar dicho modelo a todos los datos disponibles de test. En la figura 5.8 mostramos la varianza de los diez modelos para los cuatro clasificadores considerados. Como la diferencia entre número de neuronas de hombres y de mujeres es muy similar existe poca varianza en los modelos producida por el balanceo de los datos.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 969   0 |   a = Male
  11 958 |   b = Female

Correctly Classified Instances        1927               99.4324 %
Incorrectly Classified Instances        11                0.5676 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1011    1 |    a = Male
    0  969 |    b = Female

Correctly Classified Instances        1980               99.9495 %
Incorrectly Classified Instances         1                0.0505 %
```

(b) Matriz de confusión del total de datos

Figura 5.7: Matrices de confusión del clasificador SVM del género para la especie humana

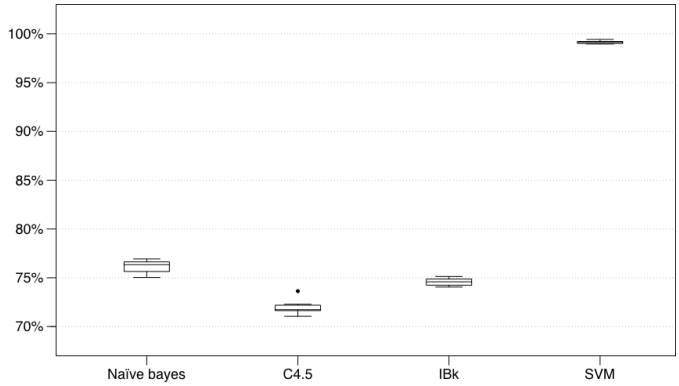


Figura 5.8: Varianza de cada clasificador para los 10 modelos de clasificación del género para la especie humana

- **Especie Ratas**: Para la especie de las ratas hemos obtenido modelos entrenando los clasificadores con todas las neuronas juntas y separando las células principales. Los mejores resultados en ambos casos han sido para el clasificador máquina de vectores soporte. Todos los modelos han obtenido resultados de precisión similares con selección de atributos wrapper como se puede observar en la figura 5.9, un poco peores para el árbol de clasificación C4.5 aunque este último utiliza 2 atributos para la clasificación de todas las neuronas (figura 5.10) frente a 17 de la red bayesiana naïve, 10 del clasificador IB1, y 4 de SVM.

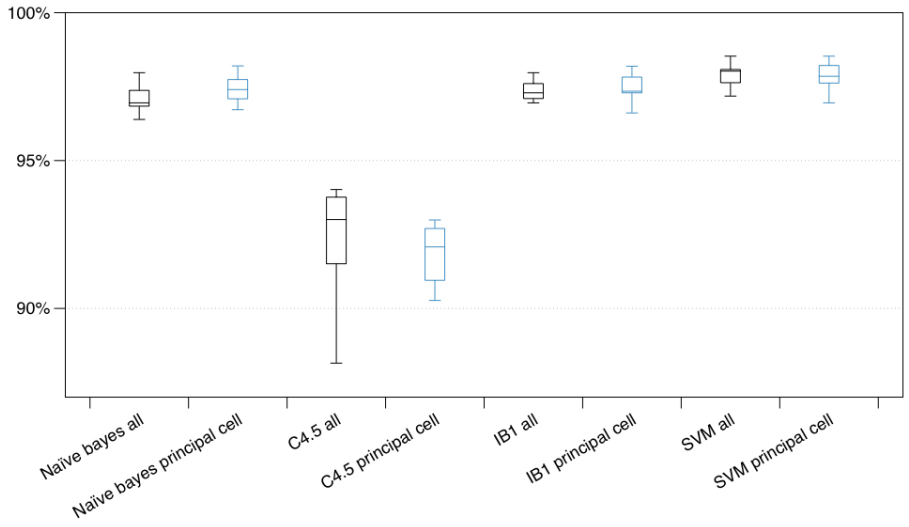


Figura 5.9: Varianza de cada clasificador para los 10 modelos de clasificación del género para la especie de las ratas

En la tabla 5.4 presentamos los atributos empleados en la clasificación del árbol C4.5: el tipo máximo y el diámetro mínimo de la última bifurcación antes del terminal final. Estos atributos no parecen estar relacionados con la literatura revisada.

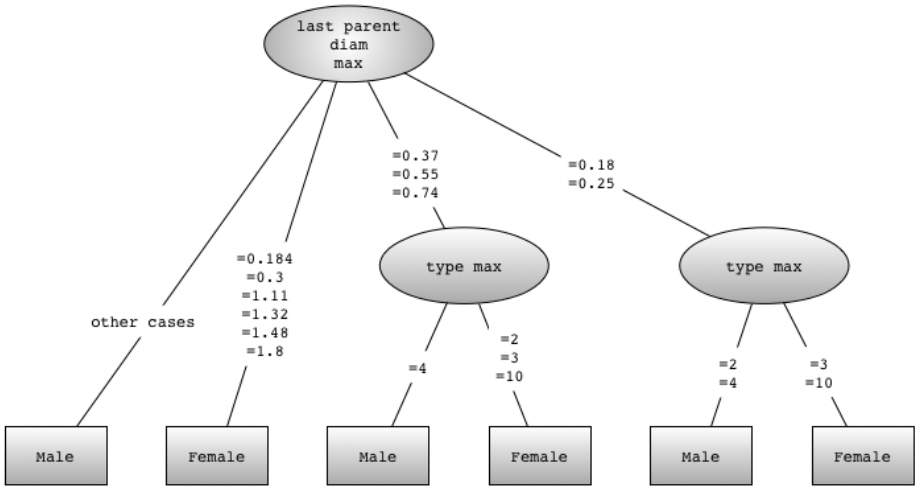


Figura 5.10: Árbol de clasificación C4.5 del género para la especie de las ratas

Número de Atributo	Nombre de Atributo	Medida
95	type	max
168	last parent diam	min

Tabla 5.4: Atributos obtenidos con selección de atributos wrapper para el árbol de clasificación C4.5 del género para la especie de las ratas

En la figura 5.11 tenemos la matriz de confusión del modelo y los resultados de aplicar el modelo seleccionado al conjunto completo de neuronas disponibles.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 401  42 |   a = Male
  11 432 |   b = Female

Correctly Classified Instances         833               94.0181 %
Incorrectly Classified Instances        53                5.9819 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1696  119 |   a = Male
   10  433 |   b = Female

Correctly Classified Instances        2129               94.287  %
Incorrectly Classified Instances       129                5.713  %
```

(b) Matriz de confusión del total de datos

Figura 5.11: Matrices de confusión del árbol de clasificación C4.5 del género para la especie de las ratas

En las células principales aunque hemos obtenido resultados similares en la precisión de los modelos (figura 5.9 (b)), el número y el tipo de atributos también es diferente. El clasificador máquina de vectores soporte ofrece los mejores resultados utilizando 4 atributos frente a 22 de la red bayesiana naïve, 17 del clasificador IB1 y 5 del árbol de clasificación C4.5. Los atributos seleccionados (tabla 5.5) son: la media del número de compartimentos que salen del soma (*n stem*), la media del número de bifurcaciones de

la neurona (*n bifs*), el tipo máximo (*type*) y la desviación estándar de la relación entre la distancia euclídea y la distancia de los compartimentos (*fractal dim*) que coinciden con los utilizados por el mismo clasificador SVM para el conjunto completo de neuronas y con la literatura, todos ellos relacionados con la longitud y la densidad dendrítica. Aunque como en los humanos ningún atributo relaciona el tamaño del soma con la diferenciación por géneros.

Número de Atributo	Nombre de Atributo	Medida
45	n stem	avg
46	n bifs	avg
95	type	max
215	fractal dim	sd

Tabla 5.5: Atributos obtenidos con selección de atributos wrapper para la clasificación SVM del género para la especie de las ratas

En la figura 5.12 tenemos las matrices de confusión del clasificador máquina de vectores soporte para la clasificación del género de las células principales de las ratas con selección de atributos wrapper.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 435   7 |   a = Male
   8 434 |   b = Female

Correctly Classified Instances         869               98.3032 %
Incorrectly Classified Instances        15                1.6968 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1369   22 |    a = Male
    5  437 |    b = Female

Correctly Classified Instances        1806               98.527  %
Incorrectly Classified Instances        27                1.473  %
```

(b) Matriz de confusión del total de datos

Figura 5.12: Matrices de confusión del clasificador SVM del género para las células principales de la especie de las ratas

- **Todas las especies**: Si observamos la distribución por géneros de las distintas especies en la figura 5.6 (b), el grupo formado por los axones únicamente tiene datos para la rata macho y el grillo hembra, por tanto lo más probable es que al clasificar se estén diferenciando las especies en lugar del género. Puede ocurrir algo similar con las interneuronas por lo que descartaremos ambas pruebas y estudiaremos el caso de las células principales por ser el más diversificado. En este conjunto de neuronas a diferencia del que contiene todas las células no aparecen ni terminales axónicos ni interneuronas lo que restará ruido a los resultados.

  Los mejores resultados se han obtenido para una de las iteraciones del clasificador IB1. En la figura 5.13 mostramos la varianza de las 10 iteraciones para los 4 modelos. Vemos en la figura 5.14 los datos del modelo y los resultados de aplicarlo al conjunto completo de neuronas. Este último resultado es muy bueno porque al entrenar el clasificador sólo utilizamos 181 neuronas por cada clase eliminando diversidad de especies, y el modelo se

ha probado sobre 3242 neuronas de machos, 1793 de hembra y 181 de hermafrodita con resultados de acierto del 85,08 %. En la tabla 5.6 aparecen los atributos utilizados: el número total de terminaciones, el número total de segmentos de las terminaciones finales, el número de dendritas que salen del soma en media, (estos tres atributos confirman los estudios mencionados que apuntan a las diferencias en la arborización y densidad dendrítica entre géneros), el tipo máximo y la media, la media y el mínimo de la relación entre la distancia euclídea y la distancia de los compartimentos (este atributo está relacionado con el tamaño y longitud de las dendritas), el ángulo máximo de bifurcación, el mínimo y la desviación estándar en la asimetría del árbol dendrítico, atributos que no habían sido considerados en la literatura en la diferencia entre géneros.

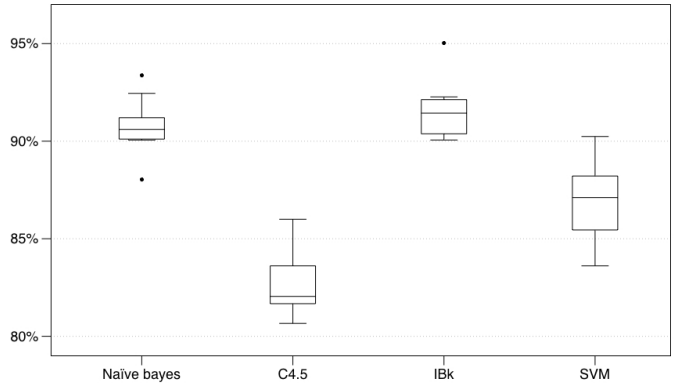


Figura 5.13: Varianza de cada clasificador para los 10 modelos de clasificación del género para todas las especies

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 162  19   0 |   a = Male
   6 174   1 |   b = Female
   1   0 180 |   c = Hermaphrodite

Correctly Classified Instances         516               95.0276 %
Incorrectly Classified Instances        27                4.9724 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 2569  635   38 |    a = Male
  103 1688    2 |    b = Female
    0    0  181 |    c = Hermaphrodite

Correctly Classified Instances        4438               85.0844 %
Incorrectly Classified Instances       778               14.9156 %
```

(b) Matriz de confusión del total de datos

Figura 5.14: Matrices de confusión del clasificador IB1 del género para todas las especies

Número de Atributo	Nombre de Atributo	Medida
19	terminal degree	total
20	terminal segment	total
45	n stem	avg
52, 95	type	avg, max
86, 172	fractal dim	avg, min
124	bif torque remote	max
143	section area	min
150	taper 1	min
157, 200	partition asymetry	min, sd

Tabla 5.6: Atributos obtenidos con selección de atributos wrapper para la clasificación IB1 por género de todas las especies

Los modelos obtenidos son capaces de diferenciar el género independientemente de la especie lo que parece indicar similitudes en las neuronas del mismo sexo para las distintas especies. En ningún caso se han obtenido atributos relacionados con el tamaño del soma, pero sí atributos relacionados con la densidad y la longitud dendrítica.

5.3. Clasificación de la edad

La estructura básica del cerebro se desarrolla y organiza antes de nacer. Sin embargo en el nacimiento no está completa sino que las experiencias tienen un papel muy importante en la evolución de las conexiones. Se sabe que a medida que aumenta la edad disminuye el número de neuronas, pero Burke y Barnes (2006) creen que estos cambios no son los responsables de la pérdida cognitiva producida por la edad, sino que se debe a los cambios en la morfología de las neuronas. En la figura 5.15 publicada en su estudio vemos el efecto del envejecimiento neuronal en la morfología.

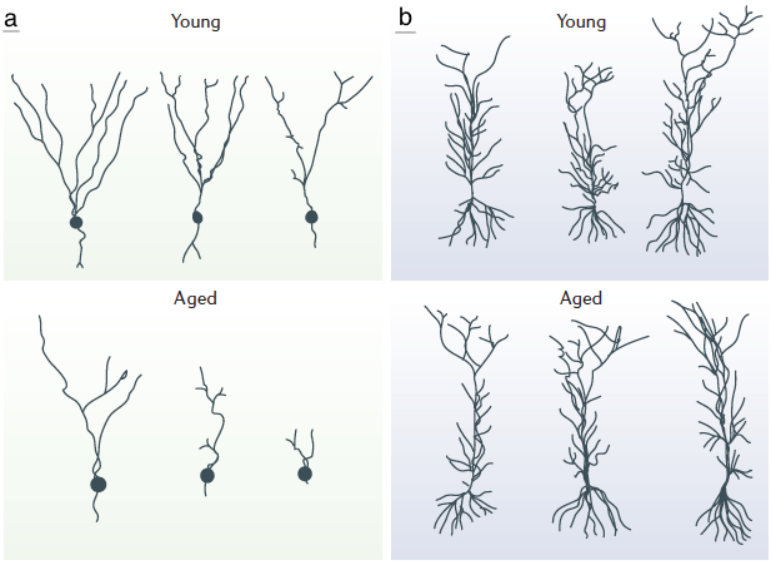


Figura 5.15: Burke y Barnes (2006): Evolución morfológica de las neuronas con la edad. (a) Neurona humana; (b) neurona de rata

Grill y Riddle (2002) han cuantificado la extensión dendrítica y la geometría de las neuronas piramidales del córtex medio frontal en ratas comprobando que se reduce la densidad y la extensión del árbol dendrítico con la edad. Markham y Juraska (2002) llegan a las mismas conclusiones, pero añaden que la disminución mencionada es más acentuada en machos. Estas diferencias no son exclusivas de las ratas y también se han estudiado en humanos. De Brabander *et al.* (1998) encuentran diferencias dependiendo de la capa del córtex prefrontal: una de las capas no presentaba cambios con la edad mientras que la otra sí. Uylings y De Brabander (2002) han publicado una revisión del estado del arte sobre los estudios referentes al tamaño de los árboles dendríticos con respecto de la edad y la demencia. Schuldiner (2014) evalúa el cambio dendrítico en el desarrollo de la mosca drosophila.

En la figura 5.16 tenemos la distribución de neuronas obtenida del NeuroMorpho por cantidad y grupo de edad que hemos utilizado en nuestras pruebas. En la mosca drosophila el desarrollo embrionario o embrión tiene lugar en el huevo, la larva es la fase posterior al huevo y anterior a la fase pupa. En los humanos neonato es un bebe recién nacido de un mes o menos, mientras en las ratas y ratones se trata de un animal de hasta 21 días de edad. Los humanos somos jóvenes hasta los 20 años aproximadamente que pasamos a la edad adulta, y viejos a partir de los 65. En cambio las ratas y ratones jóvenes tienen entre 3-8 semanas, pasan a ser adultos a partir de los 2 meses y viejos a partir de los 14 meses de edad. Los resultados obtenidos para los distintos clasificadores están en las figuras B.9, B.10, B.11 y B.12 del anexo. Al ver los resultados observamos que al trabajar con las interneuronas en las especies rata y ratón mejora la precisión, especialmente en los ratones, donde estudiaremos el grupo de interneuronas por separado. Resumimos a continuación los atributos obtenidos para cada especie considerada y las matrices de confusión para los mejores modelos obtenidos.

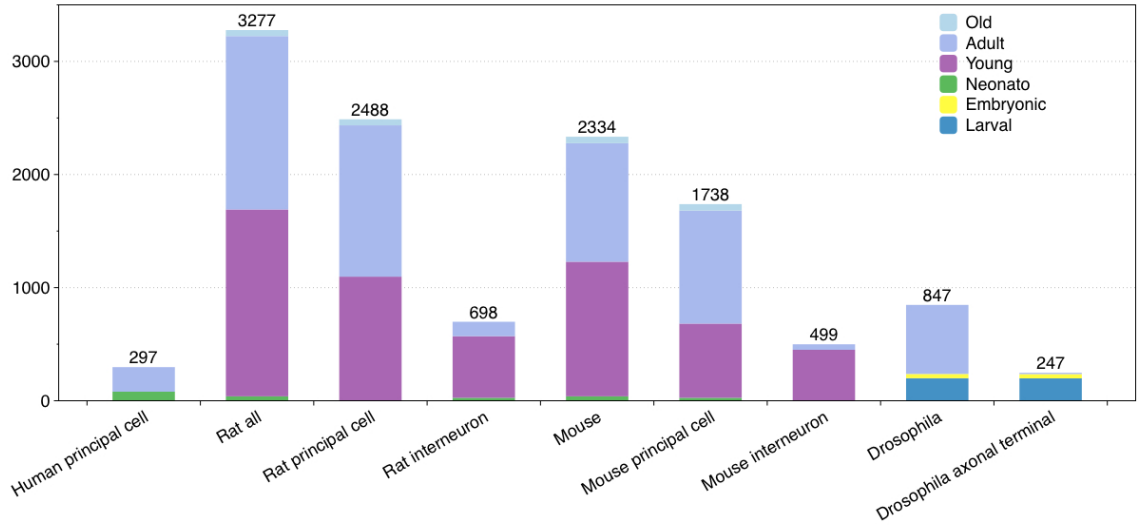
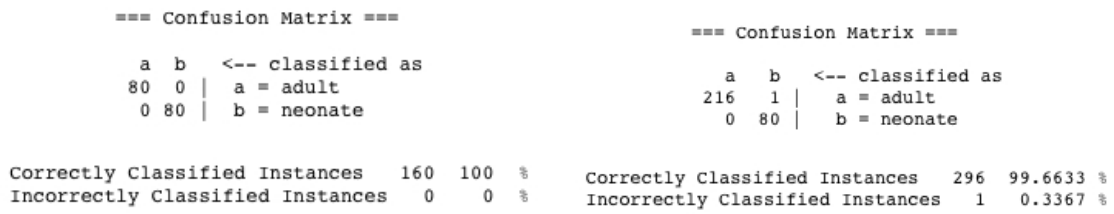


Figura 5.16: Neuronas de NeuroMorpho por edad (vertical), especie y tipo (horizontal)

- **Especie humana**: las neuronas disponibles pertenecen a las clases neonato y adulto. Se han obtenido muy buenos resultados con porcentajes de acierto cercanos al 100 % para todos los clasificadores, lo que indica grandes diferencias en la morfología neuronal de ambos. Al profundizar en los atributos obtenidos con la selección de atributos wrapper mostrados en la tabla 5.7 encontramos el valor mínimo de diámetro de los compartimentos y el máximo y la suma total del diámetro de la última bifurcación antes del terminal final. Ningún atributo coincide con la literatura que hace referencia a la densidad de los árboles dendríticos. En la figura 5.17 tenemos la matriz de confusión del modelo y el resultado de aplicar el modelo al conjunto completo de neuronas disponibles para la edad en humanos.

Número de Atributo	Nombre de Atributo	Medida
39, 125	last parent diam	total, max
139	diameter	min

Tabla 5.7: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve de clasificación por edades para la especie humana



(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.17: Matrices de confusión del clasificador bayesiano naïve por edades para la especie humana

En la figura 5.18(a) tenemos parte de los cálculos de las probabilidades realizado por el clasificador bayesiano naïve, mientras en 5.18(b) aparecen las frecuencias. La probabilidad de pertenecer a la clase adulto o neonato es $P = 0,5$ porque las clases se han balanceado. Para los 3 atributos tenemos la probabilidad condicionada dada la clase $P(X_i|C)$, como se asume independencia condicional entre los atributos se calcula la probabilidad como por ejemplo, para el primer valor del primer atributo de la clase adulto: $P_1(X_1 = 7.22|C = adult) = \frac{2}{292} = 0,006849$.

```
Class adult: P(C) = 0.5

Attribute last_parent_diam_total
7.22          9.25          5.21          5.88          4.03  …
0.00684932    0.00342466    0.01027397    0.00342466    0.0068493 …

Attribute last_parent_diam_max
1.51          3.19          2.35          4.03          2.52 …
0.01796407    0.01796407    0.01197605    0.01796407    0.01796407 …

Attribute diameter_min
0.34          1.01          0.84          0.67          0.5 …
0.01960784    0.04901961    0.05882353    0.03921569    0.03921569 …

Class neonate: P(C) = 0.5

Attribute last_parent_diam_total
7.22          9.25          5.21          5.88          4.03 …
0.00342466    0.00342466    0.00342466    0.00684932    0.00342466 …

Attribute last_parent_diam_max
1.51          3.19          2.35          4.03          2.52 …
0.00598802    0.00598802    0.00598802    0.00598802    0.01197605 …

Attribute diameter_min
0.34          1.01          0.84          0.67          0.5 …
0.00980392    0.00980392    0.03921569    0.00980392    0.00980392 …
```

(a) Probabilidades

```
Class adult: Prior probability = 0.5

last_parent_diam_total:  Discrete Estimator. Counts =  2 1 3 1 2 ... (Total = 292)
last_parent_diam_max:  Discrete Estimator. Counts =  3 3 2 3 3 ... (Total = 167)
diameter_min:  Discrete Estimator. Counts =  2 5 6 4 4 ... (Total = 102)

Class neonate: Prior probability = 0.5

last_parent_diam_total:  Discrete Estimator. Counts =  1 1 1 2 1 ... (Total = 292)
last_parent_diam_max:  Discrete Estimator. Counts =  1 1 1 1 2 ... (Total = 167)
diameter_min:  Discrete Estimator. Counts =  1 1 4 1 1 ... (Total = 102)
```

(b) Frecuencias

Figura 5.18: Modelo bayesiano naïve de clasificación por edades para las células principales de la especia humana. Se muestran los 5 primeros valores de cada atributo junto con sus probabilidades y frecuencias

En la figura 5.19 mostramos la varianza de los 10 modelos de cada clasificador.

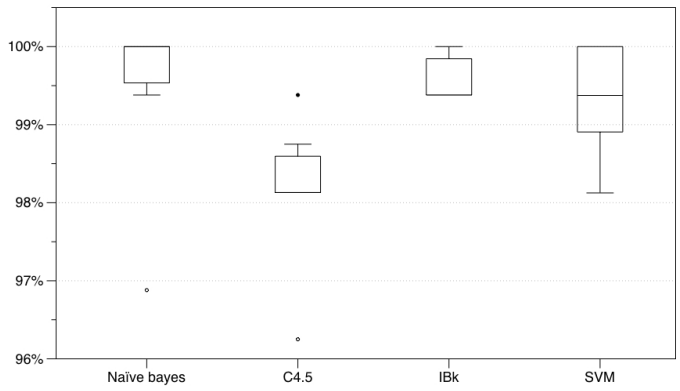


Figura 5.19: Varianza de cada clasificador para los 10 modelos de clasificación por edades para la especie humana

- **Especie rata**: las clases que tenemos disponibles para esta especie son neonato, joven, adulto y viejo (figura 5.16) para todas las neuronas. Hemos obtenido buenos resultados con los modelos IB1 y naïve Bayes, pero cuando aplicamos ambos a la totalidad de datos disponibles (figuras 5.20 (b) y (d)) obtenemos 72,65 % de aciertos del modelo IB1 frente a 70,85 % en el modelo bayesiano. Por este motivo analizaremos los atributos del primero. Las matrices de confusión de los modelos pueden verse en las figuras 5.20 (a) y (c). Si evaluamos los atributos (tabla 5.8) tenemos el total y la desviación estándar de la superficie del soma, el tipo máximo, el diámetro máximo y mínimo del último compartimento antes del terminal final, el mínimo del diámetro y de la longitud de los compartimentos (este último relacionado con la densidad y longitud del árbol dendrítico como afirman las investigaciones mencionadas), y por último las relaciones entre los diámetros de los compartimentos (*diam threshold* y *rall power*). Además los atributos *last parent diam* y *diameter* se obtuvieron también en la especie humana. En la figura 5.21 mostramos la varianza de los 10 modelos de cada clasificador, donde podemos observar que los mejores resultados han sido para los clasificadores analizados.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 27  5  0  7 |  a = adult
  6 31  0  2 |  b = young
  0  0 39  0 |  c = neonate
  1  0  1 37 |  d = old

Correctly Classified Instances         134               85.8974 %
Incorrectly Classified Instances        22               14.1026 %
```

(a) Matriz de confusión del modelo bayesiano naïve

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1064  195   95  179 |    a = adult
  254 1168   75  155 |    b = young
    0    0   39    0 |    c = neonate
    1    1    0   51 |    d = old

Correctly Classified Instances        2322               70.8575 %
Incorrectly Classified Instances       955               29.1425 %
```

(b) Matriz de confusión del total de datos del modelo bayesiano naïve

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 30  2  2  5 |  a = adult
  7 29  0  3 |  b = young
  2  0 36  1 |  c = neonate
  0  1  1 37 |  d = old

Correctly Classified Instances         132               84.6154 %
Incorrectly Classified Instances        24               15.3846 %
```

(c) Matriz de confusión del modelo IB1

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1169   59   85  220 |    a = adult
  265 1123  119  145 |    b = young
    0    0   39    0 |    c = neonate
    2    1    0   50 |    d = old

Correctly Classified Instances        2381               72.6579 %
Incorrectly Classified Instances       896               27.3421 %
```

(d) Matriz de confusión del total de datos del modelo IB1

Figura 5.20: Matrices de confusión del clasificador bayesiano naïve ((a) y (b)) y del clasificador IB1 ((c) y (d)) para clasificar las edades de la especie rata

Número de Atributo	Nombre de Atributo	Medida
1, 173	soma surface	total, sd
95	type	max
125, 168	last parent diam	max, min
139	diameter	min
141	length	min
169	diam threshold	min
201	rall power	sd

Tabla 5.8: Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación de la edad de la especie rata

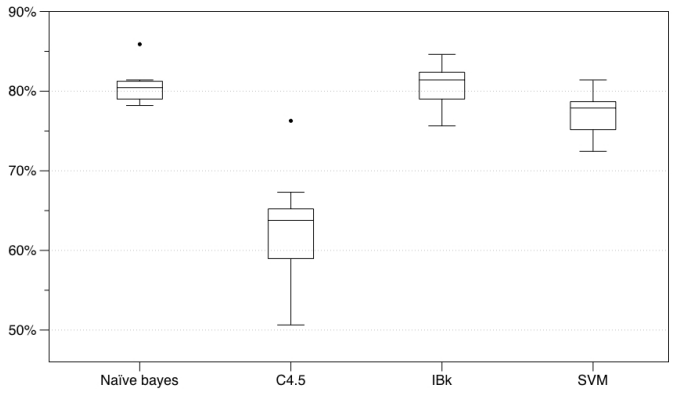
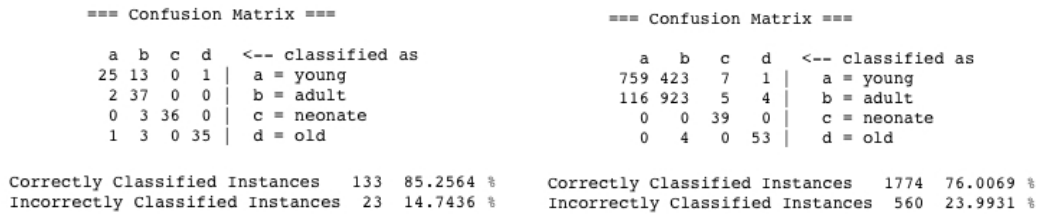


Figura 5.21: Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie rata

- **Especie ratón**: tenemos las clases neonato, joven, adulto y viejo (figura 5.16). Algunos atributos han coincidido con los obtenidos para la especie rata, hemos resaltado en negrita en la tabla 5.9 aquellos que están directamente relacionados con los de la tabla 5.8. Pero, en este caso obtenemos la media de las dendritas que salen del soma, la media de los terminales finales y la sección mínima de los compartimentos, todos ellos se refieren a la cantidad y longitud de las dendritas confirmando las diferencias en el tamaño y densidad del árbol dendrítico. En la figura 5.22 vemos la matriz de confusión del modelo máquina de vectores soporte, el mejor modelo presenta mayor sobreajuste cuando lo aplicamos a todas las neuronas disponibles. Hemos estudiado por tanto el siguiente modelo con mejor precisión y sus atributos (tabla 5.9).

Número de Atributo	Nombre de Atributo	Medida
45	n stem	avg
46	n bifs	avg
95, 138	**type**	max, min
143	section area	min
173	**soma surface**	sd

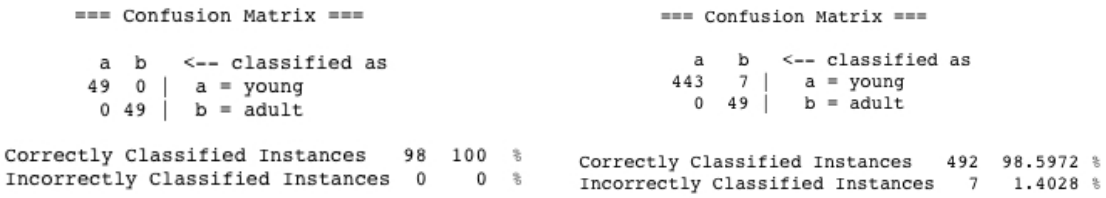
Tabla 5.9: Atributos obtenidos con selección de atributos wrapper para el modelo SVM de clasificación de la edad para la especie ratón



(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.22: Matrices de confusión del clasificador SVM de clasificación de la edad para la especie ratón

En la especie ratón se mantienen las cuatro clases (neonato, joven, adulto y viejo) al crear el grupo de células principales, en cambio en el grupo de interneuronas sólo tenemos las clases joven y adulto. A pesar de ser estas dos clases las que presentan mayores confusiones en las matrices (figura 5.22), al separar las interneuronas se produce una mejora en la precisión (figura 5.23) con sólo un atributo wrapper seleccionado. El total en el orden de las ramas, cuanto mayor es este atributo más ramas presentará la neurona. Se mantiene la diversidad que teníamos con todas las neuronas (distintas regiones del cerebro y distintos tipos de interneuronas). En la figura 5.24 donde mostramos la varianza de los 10 modelos de los cuatro clasificadores para los tres grupos.



(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.23: Matrices de confusión del clasificador SVM de clasificación de la edad para las interneuronas de la especie ratón

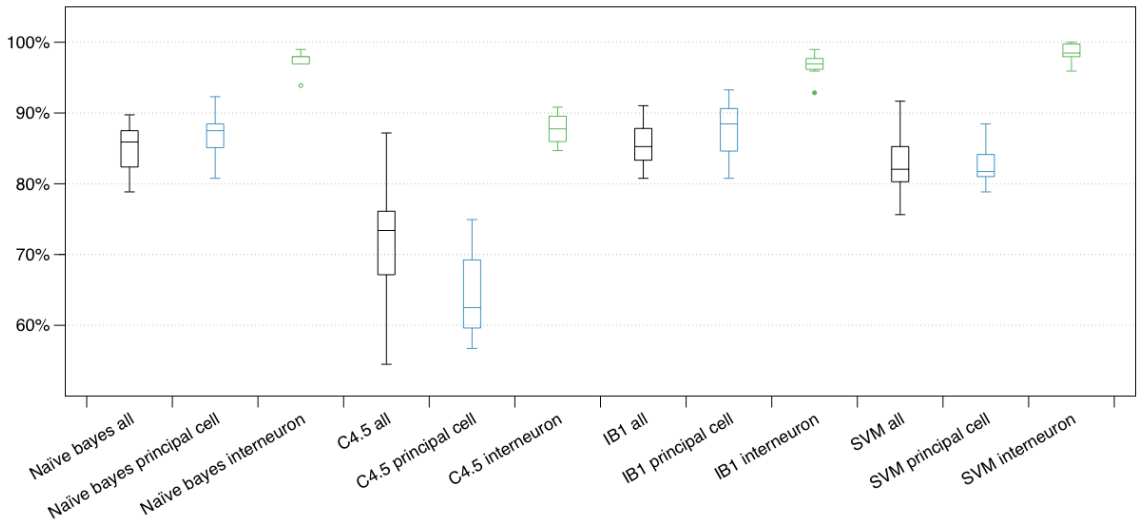


Figura 5.24: Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie ratón para los distintos grupos de neuronas

- **Especie drosophila**, las neuronas de esta especie pertenecen a las clases adulto, larva y embrión. Para este último todas las células son de tipo axón. A continuación mostramos en la figura 5.25 la varianza de los 10 modelos de cada clasificador para los distintos grupos de neuronas considerados (todas las células y terminales axónicos).

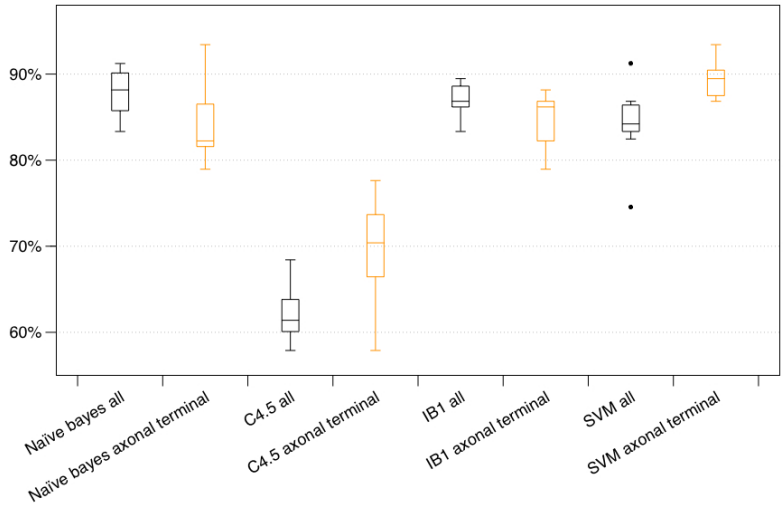
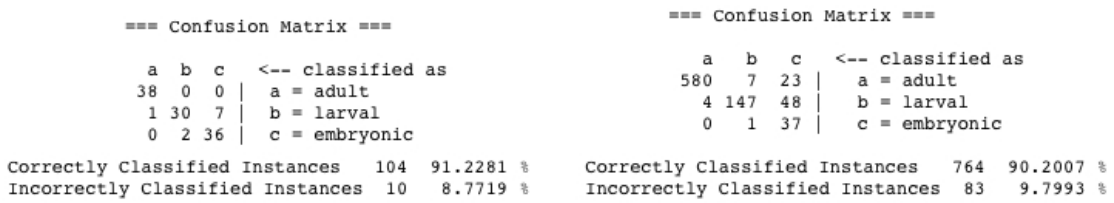


Figura 5.25: Varianza de cada clasificador para los 10 modelos de clasificación de la edad de la especie drosophila para los distintos grupos de neuronas

- **Todas las células** Para todos los clasificadores hemos obtenido resultados similares excepto para el árbol de clasificación C4.5 como vemos en la figura 5.25. Las matrices de confusión del mejor modelo y los resultados de aplicar el modelo a las neuronas disponibles aparecen en la figura 5.26. Entre los atributos seleccionados

con el método wrapper, al igual que en los casos anteriores confirman la literatura revisada y el estudio realizado por Schuldiner (2014) que señala un cambio en las dendritas y el axón conforme la mosca se desarrolla (número de bifurcaciones y de puntos terminales totales).



(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.26: Matrices de confusión del clasificador bayesiano naïve por edades para la especie drosophila

Número de Atributo	Nombre de Atributo	Medida
3	n bifs	total
5	n tips	total
52, 138	type	avg, min
104	branch order	max
126	diam threshold	max
139	diameter	min
144	volume	min
172	fractal dim	min

Tabla 5.10: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve de clasificación por edades para la especie drosophila

- **Terminal axónico**: no sólo las neuronas cambian con la edad, también los botones sinápticos o terminales axónicos como vemos en los resultados obtenidos con el modelo máquina de vectores soporte de la figura 5.27. Entre los atributos obtenidos (tabla 5.11) con selección wrapper están la superficie del soma y el número de dendritas que salen del soma. Ninguno de estos dos atributos tiene sentido, porque ambos son 0 en todas las instancias por tratarse del estudio de axones, los cuales no presentan soma (matriz de confusión en la figura 5.27 (a) y (b))). La contracción o relación entre la distancia euclídea y la distancia de la dendrita y la relación entre los diámetros en las bifurcaciones (pk) son determinantes para diferenciar las clases. Por tanto si eliminamos los dos primeros (figura 5.27 (c) y (d)) mejoramos la precisión del modelo al eliminar atributos que introducen ruido.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 37  1 |  a = larval
  4 34 |  b = embryonic

Correctly Classified Instances          71               93.4211 %
Incorrectly Classified Instances         5                6.5789 %
```

(a) Matriz de confusión del modelo con los 4 atributos de la selección wrapper

```
=== Confusion Matrix ===

   a   b   <-- classified as
 176  22 |   a = larval
   0  38 |   b = embryonic

Correctly Classified Instances         214               90.678  %
Incorrectly Classified Instances        22                9.322  %
```

(b) Matriz de confusión del total de datos con los 4 atributos de la selección wrapper

```
=== Confusion Matrix ===

  a  b   <-- classified as
 37  1 |  a = larval
  6 32 |  b = embryonic

Correctly Classified Instances          69               90.7895 %
Incorrectly Classified Instances         7                9.2105 %
```

(c) Matriz de confusión del modelo con los 2 atributos relevantes de la selección wrapper

```
=== Confusion Matrix ===

   a   b   <-- classified as
 183  15 |   a = larval
   0  38 |   b = embryonic

Correctly Classified Instances         221               93.6441 %
Incorrectly Classified Instances        15                6.3559 %
```

(d) Matriz de confusión del total de datos con los 2 atributos relevantes de la selección wrapper

Figura 5.27: Matrices de confusión del clasificador SVM de la edad para los terminales axónicos de la especie drosophila

Número de Atributo	Nombre de Atributo	Medida
1	soma surface	total
2	n stem	total
30	pk	total
110	contraction	max

Tabla 5.11: Atributos obtenidos con selección de atributos wrapper para el modelo SVM de clasificación de la edad para los terminales axónicos de la especie drosophila

5.4. Clasificación por tipo de célula

En NeuroMorpho encontramos tres grandes divisiones en el tipo de neuronas: las llamadas *células principales* que incluyen las células piramidales y en general células excitatorias, las llamadas *interneuronas* referidas a las células inhibitorias y los *terminales axónicos*. Podemos ampliar los datos en la figura 5.28 donde mostramos las neuronas descargadas de la base de datos para aquellas que superaban las 30 instancias. No existe un consenso general para nombrar o clasificar los tipos de neuronas como se describe en profundidad en DeFelipe *et al.* (2013), donde proponen una clasificación de las interneuronas GABAérgicas atendiendo a distintos criterios. Zaitsev (2013) resume las clasificaciones existentes y repasa las propiedades y la funcionalidad de una serie de interneuronas.

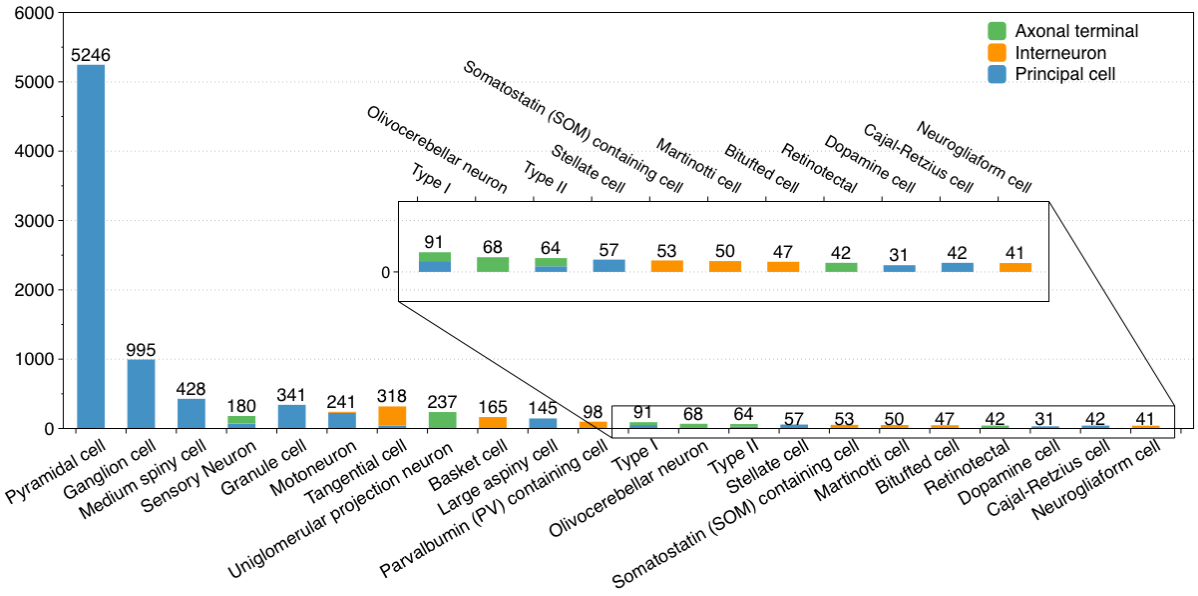


Figura 5.28: Distribución del tipo de célula en NeuroMorpho

A continuación describimos los tipos encontrados en NeuroMorpho. No siguen un patrón concreto sino que cada investigador ha asignado el tipo que ha considerado más apropiado a las neuronas que estaba tratando. Hemos creado grupos atendiendo a varios factores para darles un sentido y poder evaluar los resultados:

a) **Atendiendo a criterios morfológicos**: se centran en la idea de que la forma dicta la función. Se diferencian por la forma del axón, la ubicación de los botones sinápticos, la extensión de la neurona, etc. La forma del soma y de las dendritas puede variar mucho entre neuronas del mismo tipo y ser similar entre neuronas de distinto tipo.

- Ramón y Cajal (1893) dividió las neuronas en dos grandes grupos:
  a) *Type I*: poseen axones largos que proyectan fuera de la materia gris.
  b) *Type II*: poseen axones cortos que arborizan cerca de la célula y no salen de la materia gris.
- Distintos investigadores en su intento por llegar a un consenso (Nieuwenhuys *et al.* (2007), DeFelipe *et al.* (2013) y Zaitsev (2013)) consideran los siguientes tipos de neuronas (podemos ver dibujadas algunas de ellas en la figura 5.29):
  a) *Pyramidal cell*: poseen un axón largo y envían señales a otras regiones fuera de su área local. La forma de su soma es piramidal y presentan una gran dendrita apical y múltiples dendritas basales (Spruston (2008)).
  b) *Granule cell*: presentan varias dendritas basales cortas son extremadamente pequeñas incluido su soma y se diferencian entre ellas dependiendo de la zona en la que se encuentren, por ejemplo las del bulbo olfatorio no presentan axón.
  c) *Medium spiny cell*: presentan gran cantidad de espinas en sus dendritas. Son neuronas de tamaño medio con largos y extensos árboles dendríticos que ramifican en todas direcciones. Podemos verlas en la figura.
  d) *Large aspiny cell*: poseen un cuerpo celular largo del que radian escasas dendritas ramificadas.

e) *Cajal-Retzius*: poseen una dendrita prominente orientada horizontalmente procedente de uno de los polos del soma (Edmunds y Parnavelas (1982)).

f) *Basket cell*: sus terminales axónicos tienen forma de cesta.

g) *Martinotti cell*: son multipolares o neuronas *bitufted*, poseen un soma en forma de ovoide y el axón largo forma arborizaciones.

h) *Neurogliaform cell*: se caracterizan por un árbol dendrítico compacto y una arborización densa del axón, el soma es pequeño y esférico. Son un tipo especial de neuronas estrelladas.

i) *Ganglion cell*: son las neuronas finales de la retina. Encontramos distintos subtipos atendiendo a la morfología. Por ejemplo existen células ganglionares grandes con patrones de ramificación que irradian del núcleo y otras pequeñas similares a un arbusto (Perry *et al.* (1984)).

- Las siguientes neuronas atienden a la forma exclusivamente sin ser un nombre particular de una determinada neurona (Nieuwenhuys *et al.* (2007)).

  a) *Bitufted cell*: tienen dos dendritas principales que salen en direcciones opuestas y tras una trayectoria corta terminan en dos matas arborescentes.

  b) *Stellate cells*: poseen numerosas dendritas que irradian desde el cuerpo celular dándoles forma de estrella y las ramificaciones son poco frecuentes.

  c) *Tangential cell*: poseen axones horizontales.

b) **Atendiendo a criterios moleculares**: se utiliza la expresión de marcadores bioquímicos.

  a) *Parvalbumin containing cell (PV)*: contienen parvalbúmina que es una proteina de unión con el calcio.

  b) *Somatostatin containing cell (SOM)*: contienen el neuropéptido somatostatin.

c) **Atendiendo a los neurotransmisores**:

  a) *Dopamine cell*: son la fuente principal de dopamina en el sistema nervioso central de los mamíferos (Kalsbeek *et al.* (1992)).

d) **Atendiendo a su función**:

  a) *Motoneuron*, llevan las señales desde el cerebro y la médula espinal hasta los músculos.

  b) *Sensory neuron*, recogen información de los órganos sensoriales, los ojos, la nariz, la lengua, etc.

e) Por último dentro del grupo de los terminales axónicos encontramos: *Olivocerebellar*, *Retinotectal* y *Uniglomerular projection neuron*.

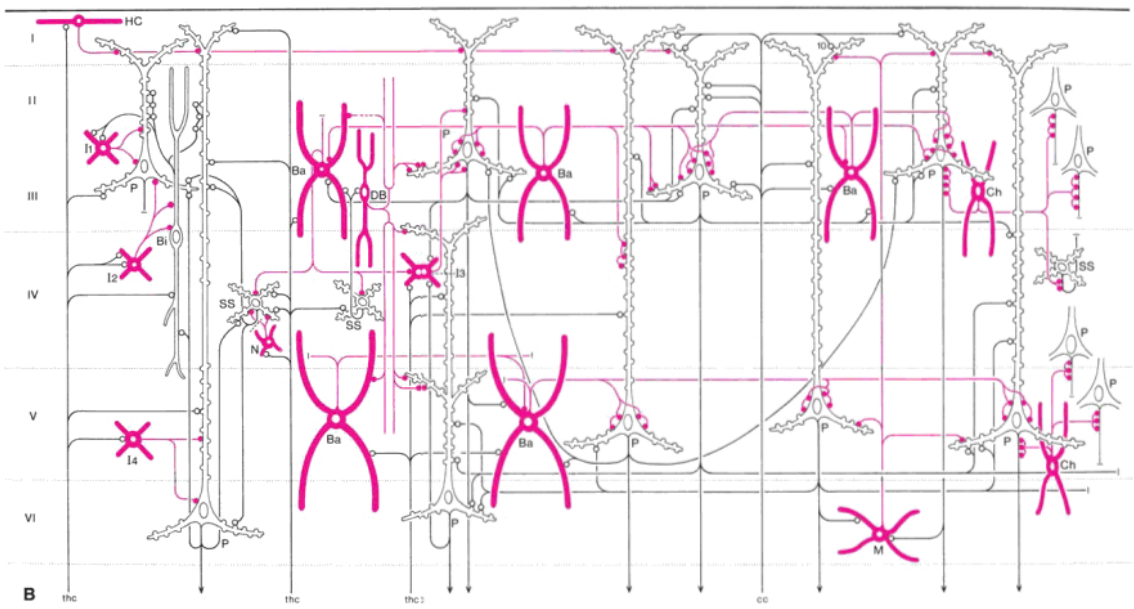
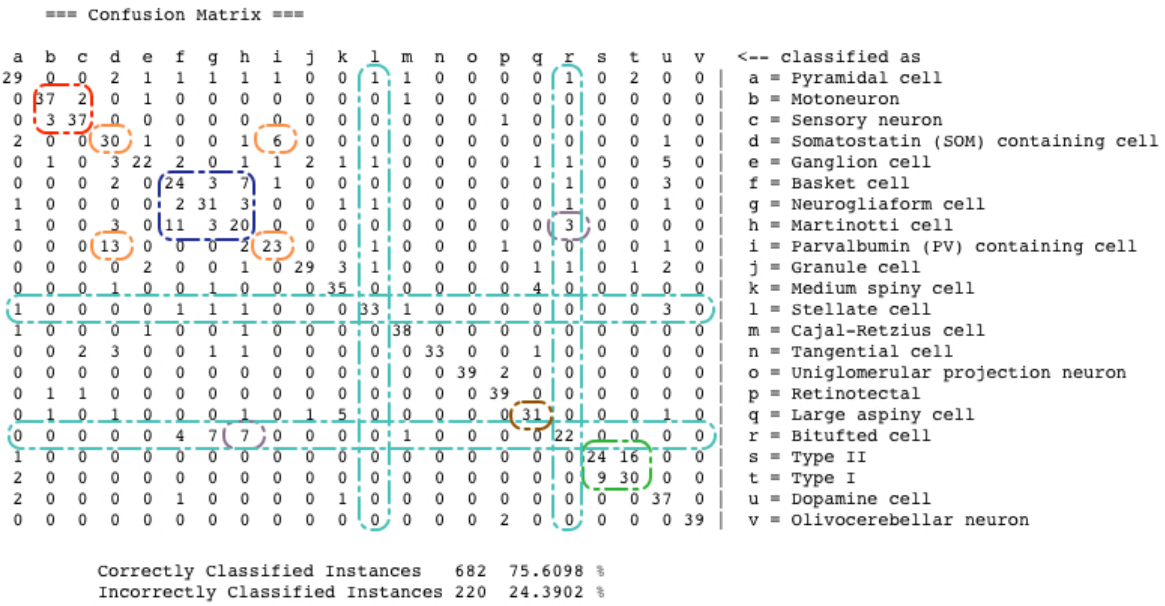


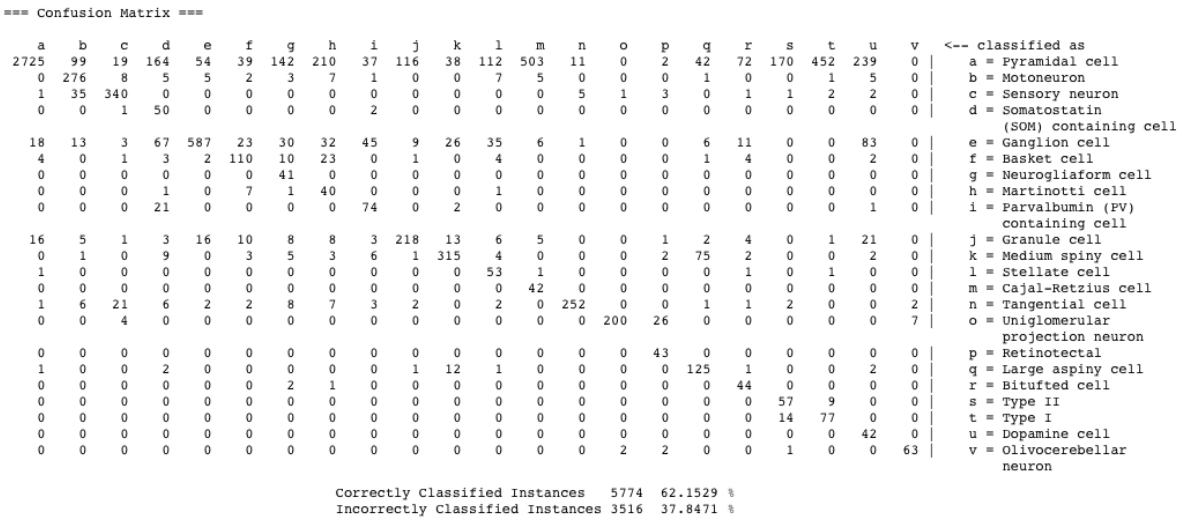
Figura 5.29: Tipos de neuronas: neuronas excitatorias en negro, neuronas inhibitorias en rosa. Ba, *basket cells*; Bi,*bipolar cell*; Ch, *Chandelier cell*; DB, *double bouquet cell*; HC: *horizontal cell of Cajal o Cajal-Retzius*; M, *Martinotti cell*; N, *neurogliaform*; P, *pyramidal neurons*; SS, *spiny stellate cells*; I_i, diferentes tipos de interneuronas. Imagen obtenida de Nieuwenhuys *et al.* (2007)

Debido a que no existe un criterio común para diferenciar los distintos tipos de neuronas hemos clasificado evaluando todas las neuronas juntas y diferenciando células principales, interneuronas y terminales axónicos. Los resultados se muestran en el anexo en las figuras B.13, B.14, B.15 y B.16.

No obstante utilizar todos los tipos descritos juntos (22 tipos de neuronas con más de 30 instancias por tipo descargadas de NeuroMorpho) no parece una buena clasificación como hemos confirmado con los resultados de la figura 5.30 para el mejor clasificador obtenido, el clasificador IB1. Se han resaltado con distintos colores los problemas que cabía esperar por juntar clases que atienden a distintos criterios y que no son excluyentes entre si. Las interneuronas Martinotti presentan en general forma *bitufted*, aparecen marcadas en morado. En naranja mostramos las neuronas que atienden a criterios moleculares, pero viendo sus resultados no parece tener sentido clasificarlas utilizando atributos morfológicos. Además, como se ha mencionado anteriormente *stellate* y *bitufted* aparentemente no son tipos de neuronas sino que serían formas de otras neuronas presentes en las clases, marcadas en azul claro. Type I y Type II marcadas en verde se confunden. Por último las motoneuronas y las neuronas sensoriales marcadas en rojo no diferencian las neuronas por morfología, sino atendiendo a su función. En azul oscuro encontramos neuronas para las que esperábamos obtener mejores resultados. Como se muestra en 5.30(b), al aplicar el modelo a todas las neuronas se acentúan los problemas mencionados.



(a) Matriz de confusión del modelo



(b) Matriz de confusión del total de datos

Figura 5.30: Matrices de confusión del clasificador IB1 por tipo de célula

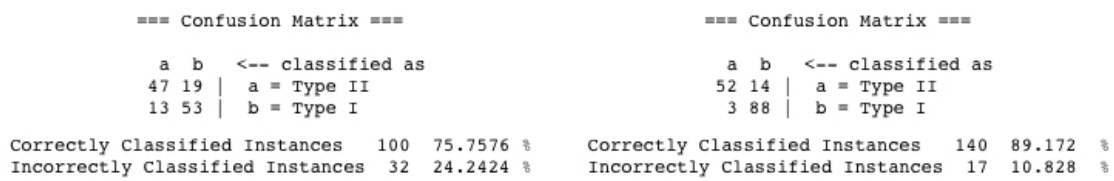
Los cuarenta atributos utilizados en la clasificación se presentan en la tabla 5.12. Entre ellos destacamos el alto, ancho y profundo de la neurona, la superficie del soma, el número de dendritas que salen del soma, el tipo de compartimentos, distintas medidas de ángulos y asimetrías, y los valores referentes a los compartimentos. Entendemos que la cantidad de atributos utilizados se debe a la dificultad para diferenciar neuronas entre grupos poco definidos y no excluyentes.

Número de Atributo	Nombre de Atributo	Medida
6, 7, 8, 50, 51	width, height, depth	total, avg
25	fragmentation	total
26	daughter ratio	total
37, 38, 123, 209	bif torque local, bif torque remote	total, max, sd
39, 125	last parent diam	total, max
40, 83, 169	diam threshold	total, avg, min
44	soma surface	avg
45	n stem	avg
52, 95, 138, 181	type	avg, max, min, sd
53, 96	diameter	avg, max
68, 111	fragmentation	avg, max
78, 121, 122	bif tilt local,bif tilt remote	avg, max
107	taper 1	max
112	daughter ratio	max
114, 200	partition asymetry	max, sd
115	rall power	max
119	bif ampl local	max
141	length	min
142	surface	min
143	section area	min
144	volume	min
172	fractal dim	min

Tabla 5.12: Atributos obtenidos con selección de atributos wrapper para el modelo IB1 para la clasificación del tipo de célula

Nos preguntamos qué resultados obtendríamos si dividiésemos las clases atendiendo a los criterios de los expertos. Para responder a esta cuestión hemos realizado pruebas con dichos criterios utilizando la red bayesiana naïve por su velocidad, obteniendo los siguientes resultados:

- Atendiendo al criterio de Ramón y Cajal: observamos una mejora respecto al modelo que incluye todos los tipos (figura 5.31) y una reducción drástica del número de atributos que permiten diferenciar entre ambas clases. Los atributos se muestran en la tabla 5.13 entre los que tenemos: el número total de dendritas que salen del soma y de bifurcaciones. Estos valores están relacionados con la densidad de los árboles dendríticos, y los diámetros del compartimento final y anterior a la última bifurcación. No aparece la profundidad o longitud de la neurona que nos permitiría diferenciar el tamaño del axón, aunque se ha confirmado la parte relacionada con la arborización que utilizaba Cajal para diferenciarlas.



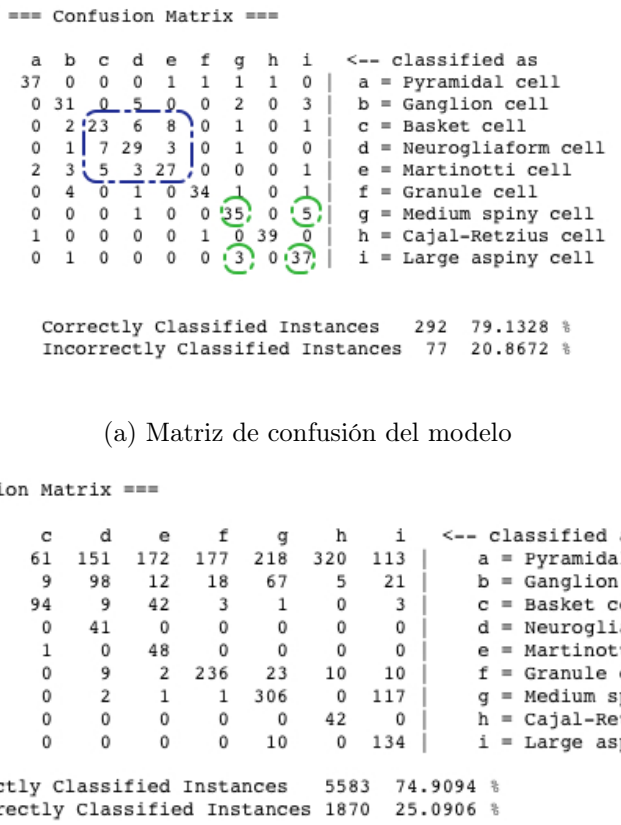
(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.31: Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los criterios de Ramón y Cajal

Número de Atributo	**Nombre de Atributo**	**Medida**
2	n stem	total
3	n bifs	total
82	last parent diam	avg
114	partition asymetry	max
151	taper 2	min
212	diam threshold	sd

Tabla 5.13: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de Ramón y Cajal

- Atendiendo a la clasificación de los investigadores: encontramos resultados similares a los obtenidos con el modelo que contiene todos los tipos de células (figura 5.32) tenemos confusiones entre las neuronas *large aspiny* y *medium spiny*, y entre las interneuronas *basket*, *neurogliaform* y *Martinotti*. Al igual que en el caso anterior, entre los atributos (tabla 5.14) aparecen el número de dendritas que salen del soma y las bifurcaciones, además del tipo máximo que nos indicará el tipo de cada compartimento y la desviación estándar en el tamaño del soma entre otros. Hemos confirmado las dificultades que existen para diferenciar ciertos tipos de interneuronas como se ha descrito en DeFelipe *et al.* (2013), donde era un grupo de expertos los que clasificaban las neuronas.



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 37  0  0  0  1  1  1  1  0 |  a = Pyramidal cell
  0 31  0  5  0  0  2  0  3 |  b = Ganglion cell
  0  2 23  6  8  0  1  0  1 |  c = Basket cell
  0  1  7 29  3  0  1  0  0 |  d = Neurogliaform cell
  2  3  5  3 27  0  0  0  1 |  e = Martinotti cell
  0  4  0  1  0 34  1  0  1 |  f = Granule cell
  0  0  0  1  0  0 35  0  5 |  g = Medium spiny cell
  1  0  0  0  0  1  0 39  0 |  h = Cajal-Retzius cell
  0  1  0  0  0  0  3  0 37 |  i = Large aspiny cell

Correctly Classified Instances         292               79.1328 %
Incorrectly Classified Instances        77               20.8672 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i   <-- classified as
 3918  116   61  151  172  177  218  320  113 |    a = Pyramidal cell
    1  764    9   98   12   18   67    5   21 |    b = Ganglion cell
    0   13   94    9   42    3    1    0    3 |    c = Basket cell
    0    0    0   41    0    0    0    0    0 |    d = Neurogliaform cell
    0    1    1    0   48    0    0    0    0 |    e = Martinotti cell
    0   51    0    9    2  236   23   10   10 |    f = Granule cell
    0    1    0    2    1    1  306    0  117 |    g = Medium spiny cell
    0    0    0    0    0    0    0   42    0 |    h = Cajal-Retzius cell
    0    1    0    0    0    0   10    0  134 |    i = Large aspiny cell

Correctly Classified Instances        5583               74.9094 %
Incorrectly Classified Instances      1870               25.0906 %
```

(b) Matriz de confusión del total de datos

Figura 5.32: Matrices de confusión del clasificador bayesiano naïve para la clasificación por tipo de célula atendiendo a los criterios de los investigadores

Número de Atributo	Nombre de Atributo	Medida
2	n stem	total
3	n bifs	total
28	partition asymetry	total
95	type	max
104	branch order	max
113	parent daughter ratio	max
115	rall power	max
116	pk	max
123	bif torque local	max
140	diameter pow	min
141	length	min
169	diam threshold	min
173	soma surface	sd

Tabla 5.14: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de los investigadores

- Atendiendo a la forma: Si seleccionamos las clases de neuronas clasificadas atendiendo a su forma, el clasificador utiliza cinco atributos que aparecen en la tabla 5.15. La precisión de los modelos es muy alta gracias a que esta división utiliza exclusivamente criterios morfológicos. Los resultados de las matrices aparecen en la figura 5.33.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 46  0  1 |  a = Stellate cell
  0 46  1 |  b = Tangential cell
  2  0 45 |  c = Bitufted cell

Correctly Classified Instances         137               97.1631 %
Incorrectly Classified Instances         4                2.8369 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
  56   0   1 |   a = Stellate cell
   4 312   2 |   b = Tangential cell
   0   0  47 |   c = Bitufted cell

Correctly Classified Instances         415               98.3412 %
Incorrectly Classified Instances         7                1.6588 %
```

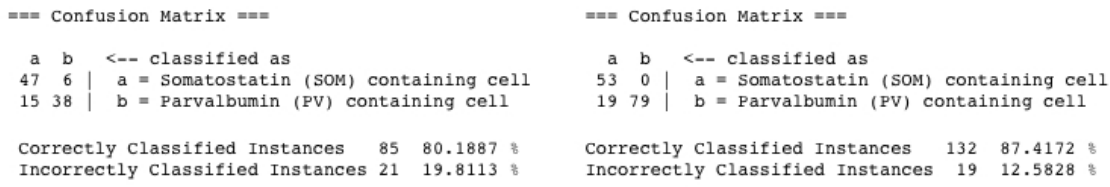
(b) Matriz de confusión del total de datos

Figura 5.33: Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los criterios de forma

Número de Atributo	Nombre de Atributo	Medida
2	n stem	total
125	last parent diam	max
139	diameter	min
169	diam threshold	min
171	helix	min

Tabla 5.15: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de forma

- Atendiendo a criterios moleculares: en los criterios moleculares no se tiene en cuenta la forma de la neurona, pero puede que la forma esté guiada por las proteinas o neuropéptidos que contiene. Tal y como podemos confirmar en la figura 5.34 los resultados son sorprendentemente buenos comparados con los que obtuvimos al clasificar estos tipos de neuronas junto con el resto. En este caso se ha utilizado el tamaño del soma, la asimetría, el tipo de compartimentos, los ángulos entre bifurcaciones y las relaciones entre los diámetros de los compartimentos (tabla 5.16).



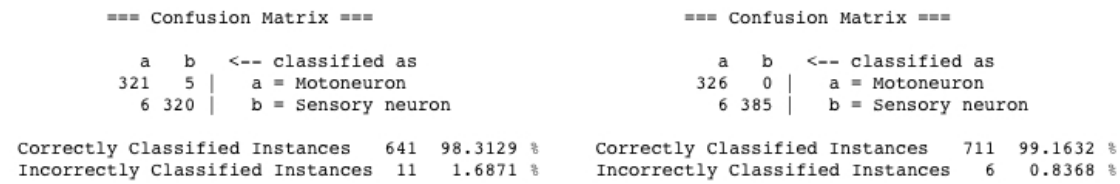
(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.34: Matrices de confusión del clasificador bayesiano naïve para la clasificación por tipo de célula atendiendo a los criterios moleculares

Número de Atributo	**Nombre de Atributo**	**Medida**
1, 44	soma surface	total, avg
71	partition asymetry	avg
82, 125	last parent diam	avg, max
95	type	max
123	bif torque local	max
124	bif torque remote	max
142	surface	min
201	rall power	sd

Tabla 5.16: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios moleculares

- Atendiendo a la función: A pesar de obtener resultados poco prometedores en estas clases cuando hemos evaluado todos los tipos de células juntos, una vez separados, las neuronas son perfectamente diferenciables morfológicamente atendiendo a la funcionalidad de la célula, como podemos comprobar en la figura 5.35. Entre los atributos de la tabla 5.17 podemos destacar el número de dendritas que salen del soma, el número de ramas, el ancho de la neurona y el tipo de compartimentos en media y desviación estándar.



(a) Matriz de confusión del modelo (b) Matriz de confusión del total de datos

Figura 5.35: Matrices de confusión del clasificador bayesiano naïve la clasificación por tipo de célula atendiendo a los criterios de funcionalidad

Número de Atributo	**Nombre de Atributo**	**Medida**
2	n stem	total
4	n branch	total
6	width	total
52, 181	type	avg, sd
70	parent daughter ratio	avg
96	diameter	max
109	branch pathlength	max
168	last parent diam	min
198	fragmentation	sd
210	bif torque local	sd
213	diam threshold	sd

Tabla 5.17: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los criterios de funcionalidad

- Atendiendo al axón: al clasificar los axones vemos que son muy diferentes entre si con resultados cercanos al 95 % de aciertos. Mostramos los resultados obtenidos con el modelo y la matriz de confusión resultante de aplicar al modelo a todos los datos disponibles en la figura 5.36. Entre los atributos mostrados en la tabla 5.18 aparecen la profundidad de la neurona, el tipo de compartimentos total y mínimo, la media del diámetro de los compartimentos, su longitud y sección, el orden de los compartimentos y algunos más. Una de las causas que ha podido favorecer los resultados son los datos que tenemos de NeuroMorpho, ya que son de distintas especies y de distintas regiones del cerebro, a excepción de las neuronas sensoriales para las que encontramos la región del sistema nervioso periférico en las especies ratón y mosca drosophila.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 40  0  0  0  0  0  0 |  a = Motoneuron
  0 40  0  0  0  0  0 |  b = Uniglomerular projection neuron
  1  0 39  0  0  0  0 |  c = Retinotectal
  0  1  0 38  0  1  0 |  d = Sensory neuron
  0  0  0  0 32  8  0 |  e = Type II
  0  0  0  0  5 35  0 |  f = Type I
  0  0  0  0  0  0 40 |  g = Olivocerebellar neuron

Correctly Classified Instances         264               94.2857 %
Incorrectly Classified Instances        16                5.7143 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  85   0   0   0   0   0   0 |   a = Motoneuron
   0 208  15   4   0   0  10 |   b = Uniglomerular projection neuron
   1   0  42   0   0   0   0 |   c = Retinotectal
   0   1   2 129   0   0   0 |   d = Sensory neuron
   0   0   0   0  40   0   0 |   e = Type II
   0   0   0   0   1  40   0 |   f = Type I
   0   0   0   0   0   0  68 |   g = Olivocerebellar neuron

Correctly Classified Instances         612               94.7368 %
Incorrectly Classified Instances        34                5.2632 %
```

(b) Matriz de confusión del total de datos

Figura 5.36: Matrices de confusión del clasificador bayesiano naïve para el criterio de clasificación del tipo de célula atendiendo a los axones

Número de Atributo	**Nombre de Atributo**	**Medida**
8	depth	total
9, 138	type	total, min
53	diameter	avg
82, 125, 168	last parent diam	avg, max, min
86, 172	fractal dim	avg, min
100, 186	section area	max, sd
104	branch order	max
110	contraction	max
126, 169	diam threshold	max, min
128	helix	max
141	length	min
160	pk classic	min

Tabla 5.18: Atributos obtenidos con selección de atributos wrapper para el modelo bayesiano naïve para la clasificación del tipo de célula atendiendo a los axones

5.5. Clasificación por región del cerebro

El cerebro y el sistema nervioso central están formados por distintas regiones y cada una de ellas está relacionada con una función concreta que conocemos gracias a las imágenes de resonancia magnética. En ellas podemos observar que aplicando ciertos estímulos exteriores se ponen en funcionamiento distintas áreas del cerebro. Hemos estudiado las neuronas por regiones y el neocórtex por separado por ser una de las partes más importantes y más estudiadas de entre las regiones cerebrales.

Entre la literatura publicada acerca de las diferencias morfológicas neuronales por regiones del cerebro, Kusicka y Schulz (1981) han estudiado las neuronas estrelladas de las ratas en tres subregiones del cerebro (neocórtex, mesoneocórtex, mesoarchicórtex) encontrando diferencias en el árbol dendrítico y las espinas de las diferentes regiones. En las mismas subregiones Schulz *et al.* (1976) realizan observaciones similares en las células piramidales, las del neocórtex y mesoneocórtex presentaban más dendritas basales mientras que las del neocórtex tenían más dendritas apicales que en las otras dos regiones. Encontramos resultados similares en publicaciones más recientes, Spruston (2008) muestran la figura 5.37 donde las neuronas piramidales son diferentes dependiendo del área cortical.

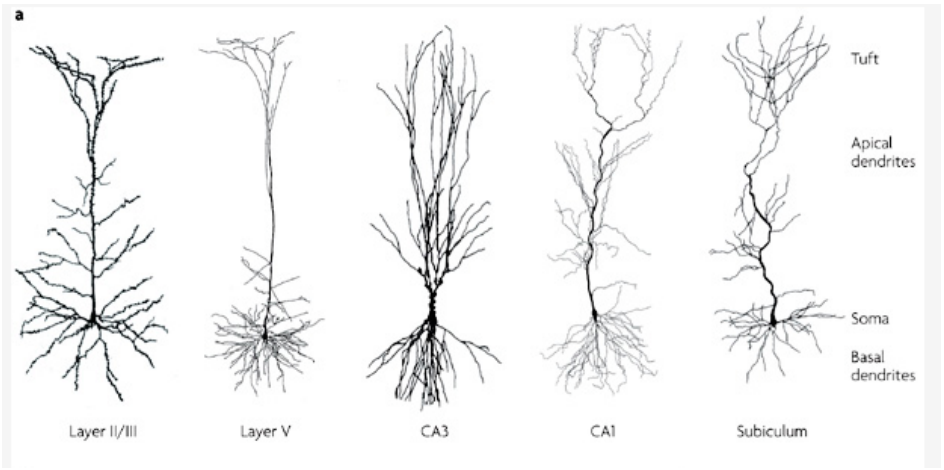


Figura 5.37: Neuronas piramidales de diferentes áreas corticales. Imagen obtenida de Spruston (2008)

5.5.1. Clasificación general

Distintas zonas del cerebro presentan los mismos tipos de neuronas. En la figura 5.38 tenemos las células disponibles para cada región del cerebro que hemos obtenido de NeuroMorpho. Como en casos anteriores hemos descartado aquellas zonas que presentan menos de 30 instancias y hemos creado distintos grupos variando el número de clases y de instancias. La media de los 10 modelos de cada clasificador se muestra en el anexo en las figuras figuras B.17, B.18, B.19 y B.20

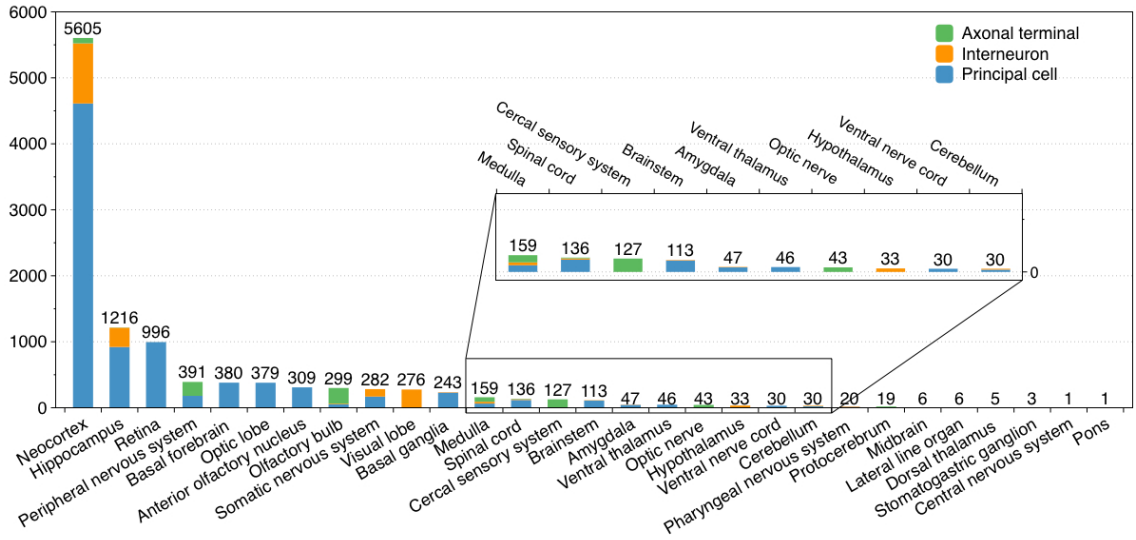


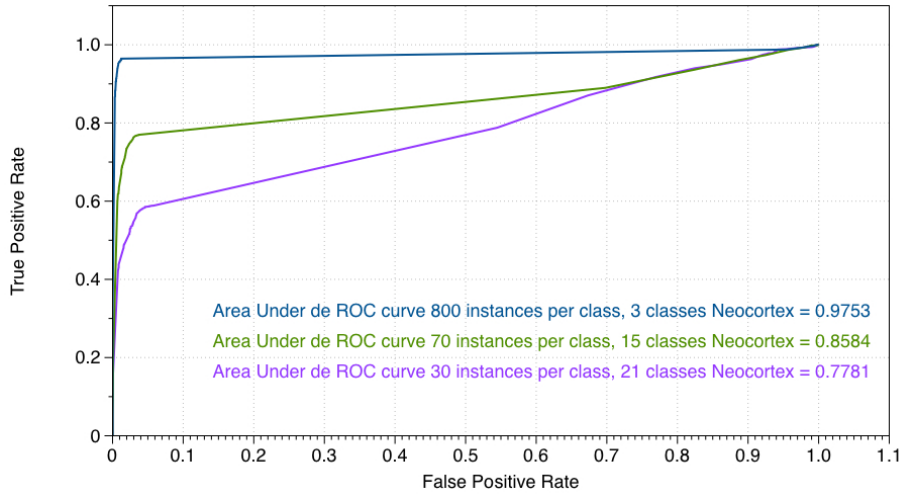
Figura 5.38: Región del cerebro

Para evaluar el efecto en la precisión que produce variar el número de clases y de instancias hemos dibujado la curva ROC de la clase neocórtex vs el resto de clases (figura 5.39(a)) para el clasificador IB1, que es para el que se han obtenido mejores resultados. A medida que reducimos el número de clases (21 clases con al menos 30 instancias, 15 clases con al menos 70 instancias y 3 clases con al menos 800 instancias) e incrementamos el número de neuronas mejoran los resultados notablemente como confirmamos con la curva ROC y con los datos de la tabla 5.19. En ella se muestra la precisión de los modelos seleccionados y los resultados de aplicar el modelo a todas las neuronas disponibles. También podemos ver en la figura 5.39(b) las curvas ROC de la clase neocórtex vs el resto de clases para el grupo de 15 clases (de neocórtex a brainstem) y al menos 70 instancias por clase para los cuatro clasificadores utilizados. Al compararlos, los mejores resultados son para la red bayesiana naïve. No obstante al aplicar la totalidad de los datos a los distintos modelos el porcentaje de aciertos es mayor con el clasificador IB1.

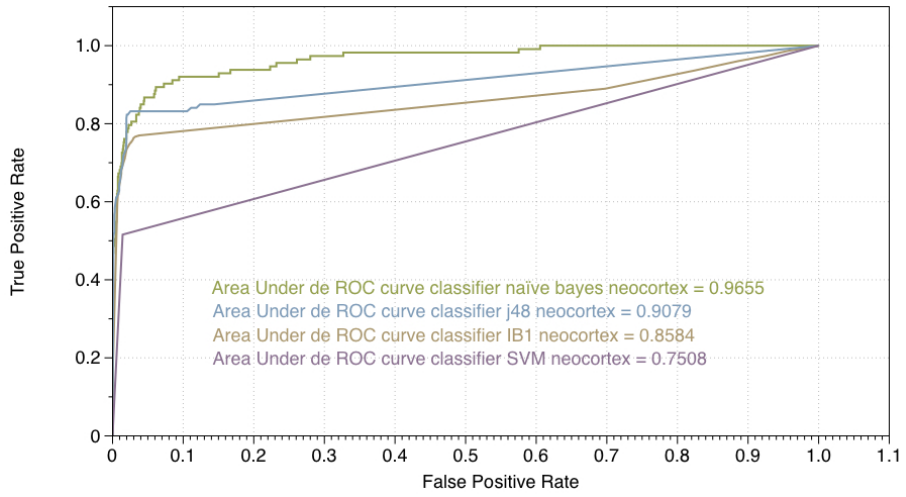
Al igual que ocurría en la clasificación por especies, no parece que 30 neuronas por clase y 21 clases sean valores para obtener modelos aceptables. Por ello hemos extraído los atributos que tenemos en la tabla 5.20 para el clasificador IB1.

21 clases 30 instancias por clase (de neocórtex a cerebellum)	**Modelo**	Correctly Classified Instances: 524	83.17 %
		Incorrectly Classified Instances: 106	16.82 %
	Total de datos	Correctly Classified Instances: 6868	61.65 %
		Incorrectly Classified Instances: 4272	38.34 %
15 clases 70 instancias por clase (de neocórtex a brainstem)	**Modelo**	Correctly Classified Instances: 1483	87.49 %
		Incorrectly Classified Instances: 212	12.50 %
	Total de datos	Correctly Classified Instances: 8568	78.52 %
		Incorrectly Classified Instances: 2343	21.47 %
3 clases 800 instancias por clase (de neocórtex a retina)	**Modelo**	Correctly Classified Instances: 2776	92.90 %
		Incorrectly Classified Instances: 212	7.1 %
	Total de datos	Correctly Classified Instances: 7449	95.29 %
		Incorrectly Classified Instances : 368	4.70 %

Tabla 5.19: Tabla de resultados del clasificador IB1 variando el número de instancias y clases



(a) Comparación de la curva ROC para la región neocórtex vs no neocórtex del clasificador IB1 con 30 instancias, 70 instancias y 800 instancias



(b) Comparación de la curva ROC para la región neocórtex vs no neocórtex de los distintos clasificadores: bayesiano naïve, árbol de clasificación C4.5, IB1 y SVM

Figura 5.39: Curva ROC

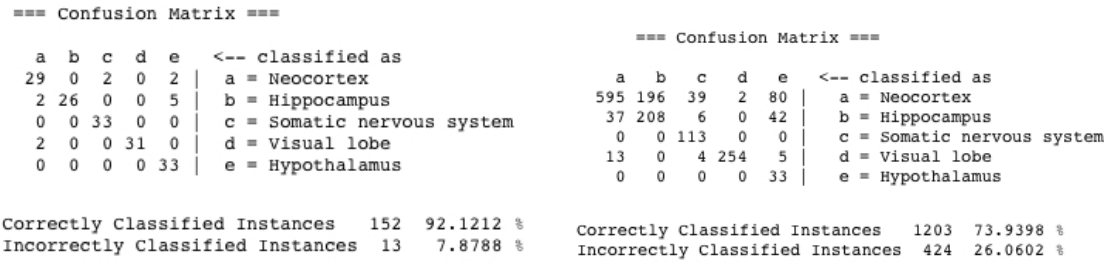
Número de Atributo	Nombre de Atributo	Medida
2,45	n stem	total, avg
8,51, 94	depth	total, avg, max
11,54, 97, 183	diameter pow	total, avg, max, sd
19	terminal degree	total
20	terminal segment	total
26	daughter ratio	total
30	pk	total
39, 125, 168	last parent diam	total,max, min
52, 95, 138, 181	type	avg, max, min, sd
73	pk	avg
96	diameter	max
107	taper 1	max
108	taper 2	max
111	fragmentation	max
113, 156	parent daughter ratio	max, min
115	rall power	max
121	bif tilt local	max
123	bif torque local	max
128	helix	max
141	length	min
142	surface	min
143	section area	min
152	branch pathlength	min
169	diam threshold	min
172	fractal dim	min
173	soma surface	sd
200	partition asymetry	sd
209	bif torque local	sd

Tabla 5.20: Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación por regiones del cerebro

- Interneuronas: si profundizamos en los datos por región del cerebro, especie y tipo de célula para las interneuronas, existe diversidad de especies para cada región del cerebro considerada (tabla 5.21). Como vemos en la figura 5.40(a) y (b) en las matrices de confusión obtenidas con el clasificador bayesiano naïve con selección de atributos wrapper para las regiones con al menos 30 instancias, ocurre lo mismo que en el caso anterior. El modelo está demasiado sobreajustada a los datos de entrenamiento y al evaluarlo con todos los datos disponibles la precisión baja al 73,93 % desde el 92,12 % del modelo. En cambio si entrenamos el clasificador con las regiones con al menos 70 instancias por clase la precisión en las matrices de confusión es notablemente mejor evitando el sobreajuste anterior (figuras 5.40(c) y (d)).

Región del cerebro	Especie
Neocortex	Cat Elephant Monkey Mouse Rat
Hippocampus	Mouse Rat
Somatic nervous system	C. Elegans
Visual lobe	Drosophila Blowfly
Hypothalamus	Mouse

Tabla 5.21: Regiones del cerebro por especies para las interneuronas

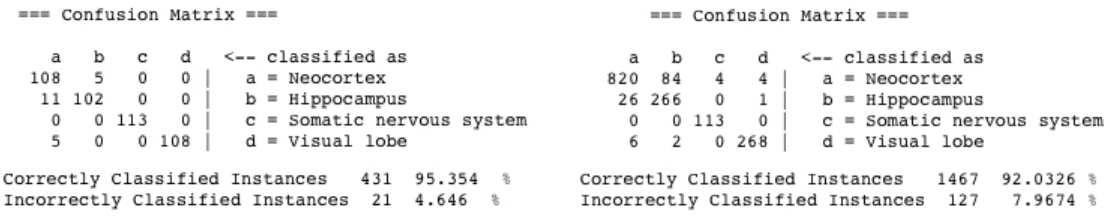


```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 29  0  2  0  2 |  a = Neocortex
  2 26  0  0  5 |  b = Hippocampus
  0  0 33  0  0 |  c = Somatic nervous system
  2  0  0 31  0 |  d = Visual lobe
  0  0  0  0 33 |  e = Hypothalamus

Correctly Classified Instances         152               92.1212 %
Incorrectly Classified Instances        13                7.8788 %
```

(a) Matriz de confusión del modelo (regiones con al menos 30 instancias por clase)

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 595 196  39   2  80 |  a = Neocortex
  37 208   6   0  42 |  b = Hippocampus
   0   0 113   0   0 |  c = Somatic nervous system
  13   0   4 254   5 |  d = Visual lobe
   0   0   0   0  33 |  e = Hypothalamus

Correctly Classified Instances        1203               73.9398 %
Incorrectly Classified Instances       424               26.0602 %
```

(b) Matriz de confusión del total de datos (regiones con al menos 30 instancias por clase)



```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 108   5   0   0 |  a = Neocortex
  11 102   0   0 |  b = Hippocampus
   0   0 113   0 |  c = Somatic nervous system
   5   0   0 108 |  d = Visual lobe

Correctly Classified Instances         431               95.354  %
Incorrectly Classified Instances        21                4.646  %
```

(c) Matriz de confusión del modelo (regiones con al menos 70 instancias por clase)

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 820  84   4   4 |  a = Neocortex
  26 266   0   1 |  b = Hippocampus
   0   0 113   0 |  c = Somatic nervous system
   6   2   0 268 |  d = Visual lobe

Correctly Classified Instances        1467               92.0326 %
Incorrectly Classified Instances       127                7.9674 %
```

(d) Matriz de confusión del total de datos (regiones con al menos 70 instancias por clase)

Figura 5.40: Matrices de confusión del clasificador bayesiano naïve para la clasificación de la región del cerebro para las interneuronas

- Terminales axónicos: al profundizar en los datos por región del cerebro y especie para el tipo terminal axónico vemos que las especies son diferentes (tabla 5.22). Pero los buenos resultados obtenidos por los distintos clasificadores (figuras B.17, B.18, B.19 y B.20), puede deberse a que se están diferenciando especies.

Región del cerebro	Especie
Olfactory bulb	Drosophila
Optic nerve	Goldfish
Peripheral nervous system	Mouse
Neocortex	Agouti
Cercal sensory system	Cricket
Medulla	Rat

Tabla 5.22: Regiones del cerebro por especies para los terminales axónicos

5.5.2. Clasificación por región del neocórtex

El córtex cerebral es el encargado de llevar a cabo las funciones superiores, tales como pensar o aprender. Las diferentes regiones del neocórtex parecen estar especializadas en distintas tareas, como por ejemplo el córtex visual o el auditivo. Tiene una estructura laminar donde cada lámina tiene neuronas de distintos tamaños y tipos (Nieuwenhuys (1994)).

Entre los tipos de neuronas que encontramos en el córtex cerebral destacan las *células piramidales* que forman aproximadamente el 70 % de las neuronas del neocórtex y son un subconjunto de las células principales. En la figura 5.41 mostramos las zonas del neocórtex y la cantidad de neuronas que tenemos disponibles para cada clase.

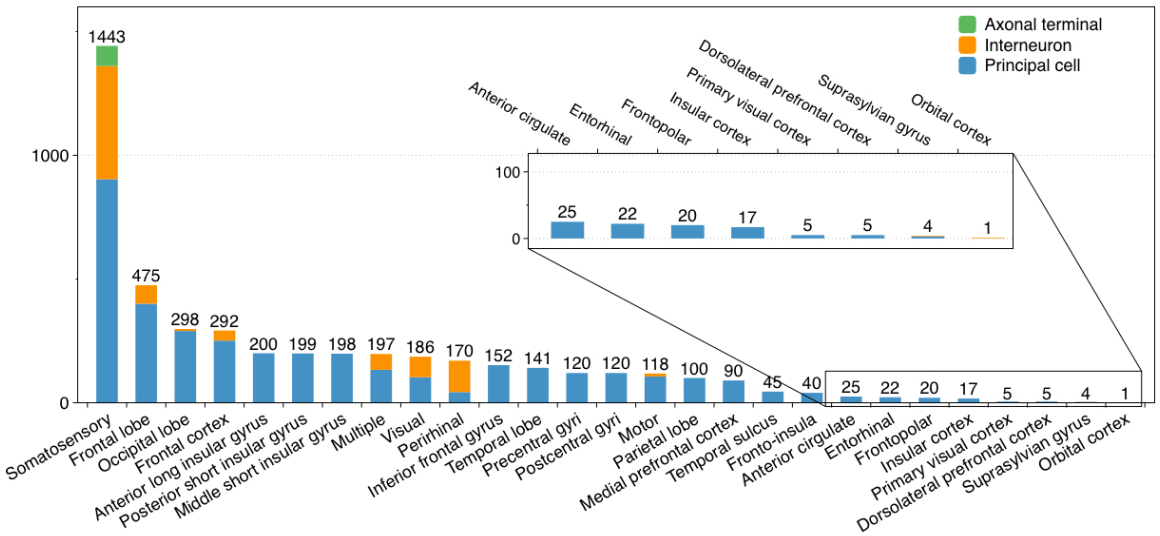


Figura 5.41: Distribución de las neuronas de NeuroMorpho por región del neocórtex

La pregunta que nos hacemos en este caso es si podemos confirmar que en general son diferenciables las neuronas entre regiones del neocórtex a pesar de que existen regiones diferentes con los mismos tipos de neuronas, es decir ¿están las neuronas del mismo tipo especializadas de alguna manera dentro de su región cortical?.

Al separar las células principales y las interneuronas, a diferencia de en las clasificaciones ya estudiadas, mejora considerablemente la precisión a pesar de que el número de clases y la cantidad de instancias se mantiene en las células principales:

- Células principales: en la figura 5.42 mostramos la matriz de confusión que pertenece al mejor modelo obtenido para las 19 regiones que presentan más de 30 instancias de neuronas. Aunque aparentemente la red bayesiana naïve con selección de atributos wrapper es la de mayor precisión, como ha ocurrido en casos anteriores, el clasificador IB1 es el que produce mejores resultados al aplicar al modelo el total de neuronas disponibles. Los atributos obtenidos aparecen en la tabla 5.23.

  Si profundizamos en los datos de las neuronas utilizadas, la mayoría son neuronas piramidales a excepción de en el lóbulo temporal que encontramos también las células Cajal-Retzius y en la zona llamada somatosensory donde además tenemos células *stellate* o estrelladas y células *tangential* o tangenciales. Podemos concluir observando los resultados de la figura 5.42 que las neuronas piramidales entre ciertas regiones son diferentes de sus homólogas, como indican Spruston (2008) y Schulz *et al.* (1976). Excepto para las zonas siguientes: subregiones que pertenecen a los lóbulos temporal, frontal, parietal y occipital marcadas en la figura en azul, las subregiones gyri precentral y postcentral, que se confunden no sólo entre ellas sino también con las anteriores, marcadas en naranja, y las subregiones insular gyrus posterior, anterior y media marcadas en verde. Además, estos resultados se mantienen en el resto de modelos para las mismas clases y también cuando incrementamos el número de instancias a al menos 70. Otro dato curioso ha sido que en este caso al aumentar el número de instancias se han obtenido peores resultados. Esto se debe a que al aumentar el número de las neuronas han desaparecido las clases *fronto-insula*, *temporal sulcus*, *perirhinal* y *medial prefrontal cortex* que son las áreas donde se obtuvieron mejores resultados.

  En la tabla 5.23 aparecen los atributos que se han utilizado en el modelo propuesto: la profundidad de la neurona, el tipo y la desviación estándar de la superficie del soma vistos anteriormente para diferenciar regiones, las longitudes y secciones de los compartimentos que se relacionan con el número de dendritas y las arborizaciones, relaciones entre los diámetros de los compartimentos y ángulos de los mismos.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s   <-- classified as
 32  5  0  0  1  0  0  0  0  0  0  0  1  0  0  0  1  0 |  a = motor
  5 22  2  0  0  0  0  0  0  1  0  0  0  3  0  5  2  0  0 |  b = somatosensory
  0  1 38  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1 |  c = fronto-insula
  0  0  0 21  3  2  2  4  4  0  0  0  0  0  0  0  4  0 |  d = temporal lobe
  0  0  0  7 17  8  1  3  2  0  0  0  0  0  0  0  0  0  2 |  e = frontal lobe
  0  0  0  6  6 12  3  8  5  0  0  0  0  0  0  0  0  0  0 |  f = parietal lobe
  0  0  0  2  2  6 12  8  8  0  0  0  0  2  0  0  0  0  0 |  g = occipital lobe
  0  0  2  5  2  4  6 18  3  0  0  0  0  0  0  0  0  0  0 |  h = precentral gyri
  0  0  0  8  3  5  3  5 15  0  1  0  0  0  0  0  0  0  0 |  i = postcentral gyri
  0  0  0  0  0  0  1  0  0 38  0  0  0  0  0  1  0  0  0 |  j = inferior frontal gyrus
  0  0  0  0  0  0  0  0  0  0 21 13  6  0  0  0  0  0  0 |  k = posterior short insular gyrus
  0  0  0  0  0  0  0  0  0  0 12 23  5  0  0  0  0  0  0 |  l = anterior long insular gyrus
  0  0  0  0  0  0  0  0  0  0 19 13  8  0  0  0  0  0  0 |  m = middle short insular gyrus
  0  2  0  0  0  0  1  0  0  0  0  0  0 33  0  3  1  0  0 |  n = multiple
  0  1  0  0  0  0  0  0  0  0  0  0  0  0 38  1  0  0  0 |  o = temporal sulcus
  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0 37  0  1  0 |  p = visual
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0 |  q = frontal cortex
  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  3 35  0 |  r = medial prefrontal cortex
  0  0  1  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0 37 |  s = perirhinal

       Correctly Classified Instances         497               65.3947 %
       Incorrectly Classified Instances       263               34.6053 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   s   <-- classified as
  98   2   0   1   0   0   0   0   0   0   0   0   0   1   0   4   0   0   0 |   a = motor
  79 431  16  12   2   0   2   1   3   1   0   0   0 102   2 121 100  25   6 |   b = somatosensory
   0   0  40   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   c = fronto-insula
   0   0   0  83   4  12   5   8  20   0   0   0   0   0   0   1   3   5   0 |   d = temporal lobe
   1   1   7  56 198  36  14  27  33   1   2   0   2   7   0   3   4   2   6 |   e = frontal lobe
   0   0   0  16   5  55   2  11  11   0   0   0   0   0   0   0   0   0   0 |   f = parietal lobe
   0   0   0  34  13  57  94  32  53   0   0   0   0   5   0   2   0   0   0 |   g = occipital lobe
   0   0   2  12   4  18   8  58  18   0   0   0   0   0   0   0   0   0   0 |   h = precentral gyri
   0   0   0  19   4  13   6  23  55   0   0   0   0   0   0   0   0   0   0 |   i = postcentral gyri
   0   0   0   0   3   0   0   0   0 148   0   0   0   0   0   0   0   0   1 |   j = inferior frontal gyrus
   0   0   0   0   0   0   0   0   0   0 118  48  33   0   0   0   0   0   0 |   k = posterior short insular gyrus
   0   0   0   0   0   0   0   0   0   0  72  96  32   0   0   0   0   0   0 |   l = anterior long insular gyrus
   0   1   0   0   0   0   0   0   0   0  50  68  78   0   0   1   0   0   0 |   m = middle short insular gyrus
   1   1   0   0   0   0   1   0   0   0   1   0   0 116   0  13   0   0   0 |   n = multiple
   0   0   0   0   0   0   0   0   0   0   0   0   0   0  45   0   0   0   0 |   o = temporal sulcus
   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0 100   2   0   0 |   p = visual
   0   0   0   0   1   0   0   0   0   0   0   0   1   5   0   2 240   2   0 |   q = frontal cortex
   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   3  85   0 |   r = medial prefrontal cortex
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  43 |   s = perirhinal

       Correctly Classified Instances        2181               60.0165 %
       Incorrectly Classified Instances      1453               39.9835 %
```

(b) Matriz de confusión del total de datos

Figura 5.42: Matrices de confusión del clasificador IB1 para la clasificación de las células principales por región del neocórtex

Número de Atributo	Nombre de Atributo	Medida
8, 51	depth	total, avg
18	branch order	total
71	partition asymetry	avg
72	rall power	avg
75	pk 2	avg
95	type	max
108, 151	taper 2	max, min
123	bif torque local	max
141	length	min
143	section area	min
152	branch pathlength	min
168, 211	last parent diam	min, sd
169	diam threshold	min
170	hillman threshold	min
173	soma surface	sd

Tabla 5.23: Atributos obtenidos con selección de atributos wrapper para el clasificador IB1 de las células principales por región del neocórtex

- Interneuronas: a diferencia de las células principales, las interneuronas presentan tipos de células muy variados. Por ejemplo para la zona somatosensory encontramos células Martinotti, *basket*, *parvalbumin*, *somatostatin* y valores sin etiquetar que desconocemos. Sin embargo los resultados han sido más precisos a pesar de compartir el mismo tipo de neuronas entre las distintas zonas. Por ejemplo encontramos neuronas de tipo Martinotti en las zonas visual, lóbulo frontal y *somatosensory* y neuronas de tipo *bitufted* en las zonas *somatosensory* y *perirhinal*. En la matriz de confusión de la figura 5.43 vemos que el clasificador diferencia el mismo tipo de neuronas de las distintas zonas perfectamente. Podríamos por tanto concluir que especialmente las interneuronas a pesar de ser del mismo tipo son capaces de diferenciarse dependiendo de la zona del córtex en la que se encuentran.
  Los resultados no son exclusivos de este modelo IB1. Al igual que en el anterior siguen todos el mismo patrón de resultados. El mejor modelo obtenido ha sido para el clasificador IB1 con selección de atributos wrapper cuyos atributos aparecen en la tabla 5.24.

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 30  4  3  3  1  0 |  a = somatosensory
  4 36  1  0  0  0 |  b = visual
  2  0 38  1  0  0 |  c = multiple
  0  0  0 40  1  0 |  d = frontal cortex
  0  0  1  9 31  0 |  e = frontal lobe
  0  0  0  0  0 41 |  f = perirhinal

Correctly Classified Instances         216       87.8049 %
Incorrectly Classified Instances        30       12.1951 %
```

(a) Matriz de confusión del modelo

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 260  76  61  31  28   2 |   a = somatosensory
   5  77   0   1   0   0 |   b = visual
   5   0  59   0   0   0 |   c = multiple
   0   0   0  41   0   0 |   d = frontal cortex
   2   0   0   7  66   0 |   e = frontal lobe
   1   0   0   0   1 125 |   f = perirhinal

Correctly Classified Instances         628       74.0566 %
Incorrectly Classified Instances       220       25.9434 %
```

(b) Matriz de confusión del total de datos

Figura 5.43: Matrices de confusión del clasificador IB1 para la clasificación de las interneuronas por región del neocórtex

Número de Atributo	**Nombre de Atributo**	**Medida**
4	n branch	total
20	terminal segment	total
39, 125	last parent diam	total, max
44, 173	soma surface	avg, sd
45	n stem	avg
95	type	max
111	fragmentation	max
112	daughter ratio	max
139	diameter	min
141	length	min
152	branch pathlength	min
159	pk	min
169	diam threshold	min

Tabla 5.24: Atributos obtenidos con selección de atributos wrapper para el modelo IB1 de clasificación de interneuronas por región del neocórtex

5.6. Comparación de los algoritmos

En la tabla 5.25 se muestra la diferencia de precisión (en %) entre los modelos sin selección de atributos y los modelos que utilizan selección CFS y wrapper. Los mejores modelos obtenidos han sido siempre al considerar la selección de atributos wrapper, mientras que la selección cfs en general empeora la precisión de los clasificadores, especialmente para el árbol de clasificación C4.5. Excepto para el clasificador máquina de vectores soporte donde utilizar todos los atributos originales es peor opción que hacer selección CFS.

	bayesiano naïve		j48		IB1		SVM	
	cfs	wrapper	cfs	wrapper	cfs	wrapper	cfs	wrapper
Especie	-2,38	+6,07	-14,83	0,38	-1,55	+5,62	+44,22	+55,38
Género	-4,03	+3,94	-4,33	+2,86	-4,17	+3,39	+42,49	+49,92
Edad	-2,79	+7,12	-6,11	+8,34	-1,11	+9,17	+24,62	+46,10
Tipo de célula	-6,76	+8,57	-25,56	+2,09	-6,74	+6,88	+28,08	+51,16
Región del cerebro	-6,14	+6,56	-20,49	+1,44	-6,49	+6,08	+39,45	+59,96
Neocórtex	-0,86	+5,24	-0,63	+1,67	+2,23	+7,95	+39,55	+62,83

Tabla 5.25: Mejora obtenida con la selección de atributos en los distintos clasificadores

En general el algoritmo j48 para la inducción de árboles de clasificación en la selección de atributos wrapper ha generado menos de la mitad de atributos que los algoritmos IB1 y la red bayesiana naïve, pero obteniendo para la mayoría de las pruebas porcentajes de acierto significativamente menores a los otros modelos, entorno a un 10 %. En cambio la máquina de vectores soporte, con menos atributos que el árbol de clasificación C4.5, ha dado buenos resultado. Los mejores modelos han sido normalmente para el algoritmo IB1, mientras que la red bayesiana naïve ha obtenido precisiones muy similares.

	todos		bayesiano naïve	j48	IB1	SVM
	sin selección	cfs	wrapper	wrapper	wrapper	wrapper
Especie	216	6,80	17,80	6,50	17,30	3,90
Género	216	3,34	13,74	3,23	11,54	4,20
Edad	216	5,33	9,22	2,11	8,22	4,33
Tipo de célula	216	6,50	15,50	6,25	19	4,25
Región del cerebro	216	5,30	18,50	6,70	18,10	2,30
Neocórtex	216	5,80	19,20	6	12,40	8

Tabla 5.26: Número medio de atributos seleccionados en los distintos clasificadores

Respecto al tiempo empleado este ha dependido de la cantidad de clases y de instancias utilizadas. El modelo más rápido, con mucha diferencia, en extraer los atributos wrapper ha sido la red bayesiana naïve que ha tardado minutos, seguida por el árbol j48 que llegaba a tardar horas similar a la máquina de vectores soporte, por último, el algoritmo IB1 que en los casos más extremos ha necesitado días para ejecutar una iteración. En cambio en los grupos con clases de pocas instancias todos los algoritmos han tenido buena respuesta en tiempo,

como por ejemplo en las pruebas de la edad o del género.

El separar los grupos formados por interneuronas o por neuronas principales no ha ofrecido resultados más precisos que aquellos grupos donde hemos trabajado con todos los tipos de células, excepto en el neocórtex como se ha comentado en la subsección 5.5.2 y en la clasificación por edades para los ratones. En cambio estudiar los terminales axónicos por separado ha proporcionado buenos resultados a la par que interesantes.

Capítulo 6

Conclusiones

6.1. Conclusiones

Hemos utilizado distintos algoritmos de aprendizaje supervisado para clasificar las neuronas obtenidas de NeuroMorpho, trabajando con todos los atributos extraídos por el software L-Measure. Se han limpiado los datos y se han realizado distintas técnicas de selección de atributos. Hemos estudiado los modelos y los atributos obtenidos para las distintas especies, géneros, edades, tipos de célula y regiones del cerebro profundizando en la división dentro del córtex cerebral.

Se han obtenido buenos resultados proporcionando modelos con precisiones altas que funcionan para todas las neuronas disponibles que no fueron utilizadas cuando se entrenó el clasificador debido al balanceo de datos. Hemos llegado a testar más de 5000 neuronas en modelos entrenados con 600 instancias. Hemos confirmado que el número de neuronas utilizadas es muy importante porque células de un mismo animal pertenecientes a una misma región del cerebro y etiquetadas por los expertos con el mismo tipo pueden ser muy diferentes entre sí.

Hemos identificado los atributos más importantes para diferenciar neuronas dependiendo de la clase, confirmando aquellos que se habían identificado bajo microscopio por los investigadores, especialmente los relacionados con el número de dendritas y las arborizaciones. También hemos encontrado atributos nuevos que no se habían estudiado y se repiten a lo largo de todas las clasificaciones, como son los relacionados con los diámetros de las dendritas. Hemos confirmado los problemas que existen al agrupar las neuronas por tipo de célula donde cada investigador asigna un tipo sin que exista consenso.

Los mejores resultados de la investigación han sido al trabajar con la mosca drosophila. Esto puede deberse a que las últimas neuronas introducidas en la base de datos NeuroMorpho correspondientes a la mosca drosophila tienen una calidad mayor.

6.2. Trabajo Futuro

Partiendo de los modelos creados se podrán etiquetar aquellas neuronas de NeuroMorpho u otras obtenidas por otros medios que aparezcan sin valor para las clases estudiadas.

La neurociencia es un campo en el que quedan aportaciones importantes por realizar. Se podrá profundizar en los datos combinando las clases estudiadas. Por ejemplo verificar los

resultados de Markham y Juraska (2002) comprobando el efecto de la edad dependiendo del género, estudiar las zonas cerebrales que presentan mayor degeneración con la edad como afirman De Brabander *et al.* (1998) o comprobar el dimorfismo sexual en distintas zonas y tipos de neuronas, etc.

Una de las partes más importantes será comprobar las diferencias morfológicas entre neuronas sanas y neuronas que presenten enfermedades degenerativas o adicciones, pudiendo evaluar qué está ocurriendo en dichos casos.

Sería importante ajustar los parámetros de los modelos para mejorar los resultados. Para ello se pueden utilizar algoritmos genéticos que encuentren los valores óptimos de los mismos. También se pueden combinar algoritmos o utilizar metaclasificadores (*ensemble*) para crear modelos más complejos.

Anexo A

Atributos L-Measure

Número del atributo	Nombre del atributo	Valor del atributo
1, 44, 87, 130, 173	soma surface	total, avg, max, min, sd
2, 45, 88, 131, 174	n stem	total, avg, max, min, sd
3, 46, 89, 132, 175	n bifs	total, avg, max, min, sd
4, 47, 90, 133, 176	n branch	total, avg, max, min, sd
5, 48, 91, 134, 177	n tips	total, avg, max, min, sd
6, 49, 92, 135, 178	width	total, avg, max, min, sd
7, 50, 93, 136, 179	height	total, avg, max, min, sd
8, 51, 94, 137, 180	depth	total, avg, max, min, sd
9, 52, 95, 138, 181	type	total, avg, max, min, sd
10, 53, 96, 139, 182	diameter	total, avg, max, min, sd
11, 54, 97, 140, 183	diameter pow	total, avg, max, min, sd
12, 55, 98, 141, 184	length	total, avg, max, min, sd
13, 56, 99, 142, 185	surface	total, avg, max, min, sd
14, 57, 100, 143, 186	section area	total, avg, max, min, sd
15, 58, 101, 144, 187	volume	total, avg, max, min, sd
16, 59, 102, 145, 188	euc distance	total, avg, max, min, sd
17, 60, 103, 146, 189	path distance	total, avg, max, min, sd
18, 61, 104, 147, 190	branch order	total, avg, max, min, sd
19, 62, 105, 148, 191	terminal degree	total, avg, max, min, sd
20, 63, 106, 149, 192	terminal segment	total, avg, max, min, sd
21, 64, 107, 150, 193	taper 1	total, avg, max, min, sd
22, 65, 108, 151, 194	taper 2	total, avg, max, min, sd
23, 66, 109, 152, 195	branch pathlengh	total, avg, max, min, sd
24, 67, 110, 153, 196	contraction	total, avg, max, min, sd
25, 68, 111, 154, 197	fragmentation	total, avg, max, min, sd
26, 69, 112, 155, 198	daughter ratio	total, avg, max, min, sd
27, 70, 113, 156, 200	parent daughter ratio	total, avg, max, min, sd
28, 71, 114, 157, 201	partition asymetry	total, avg, max, min, sd
29, 72, 115, 158, 202	rall power	total, avg, max, min, sd
30, 73, 116, 159, 203	pk	total, avg, max, min, sd
31, 74, 117, 160, 204	pk classic	total, avg, max, min, sd
32, 75, 118, 161, 205	pk 2	total, avg, max, min, sd
33, 76, 119, 162, 206	bif ampl local	total, avg, max, min, sd
34, 77, 120, 163, 207	bif ampl remote	total, avg, max, min, sd
35, 78, 121, 164, 208	bif tilt local	total, avg, max, min, sd
36, 79, 122, 165, 209	bif tilt remote	total, avg, max, min, sd
37, 80, 123, 166, 210	bif torque local	total, avg, max, min, sd
38, 81, 124, 167, 211	bif torque remote	total, avg, max, min, sd
39, 82, 125, 168, 212	last parent diam	total, avg, max, min, sd
40, 83, 126, 169, 213	diam threshold	total, avg, max, min, sd
41, 84, 127, 170, 214	hillman threshold	total, avg, max, min, sd
42, 85, 128, 171, 215	helix	total, avg, max, min, sd
43, 86, 129, 172, 216	fractal dim	total, avg, max, min, sd

Tabla A.1: Atributos del software L-Measure

Anexo B

Gráficas de resultados

B.1. Clasificación de la especie

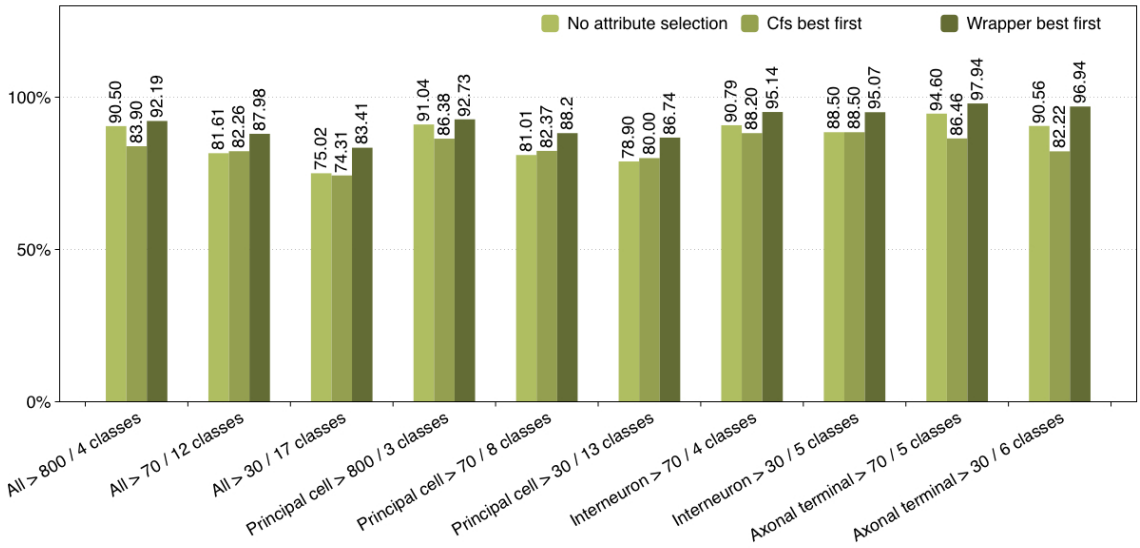


Figura B.1: Clasificación de la especie para el clasificador bayesiano naïve

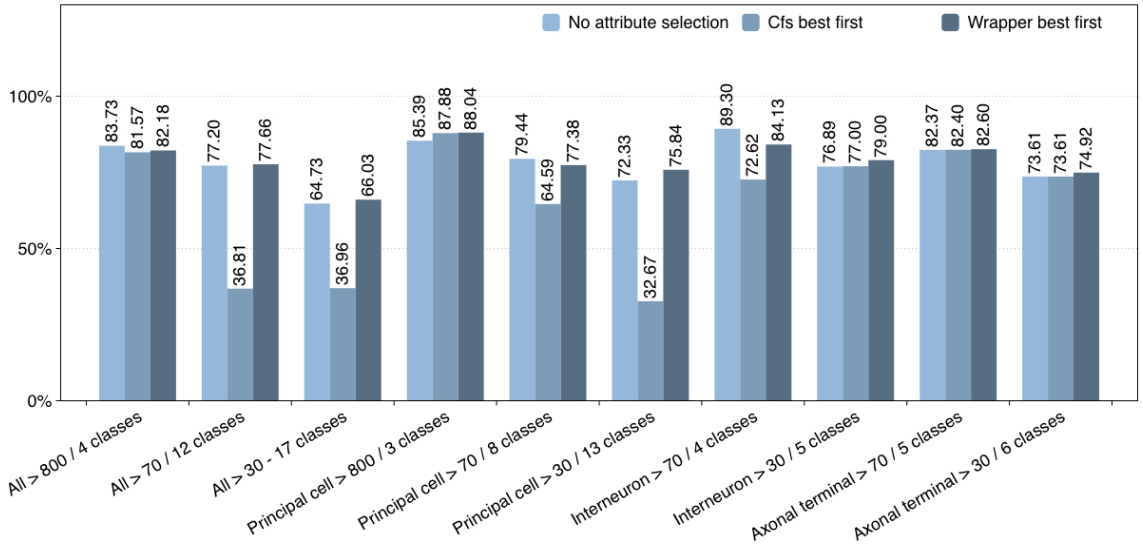


Figura B.2: Clasificación de la especie para el clasificador j48

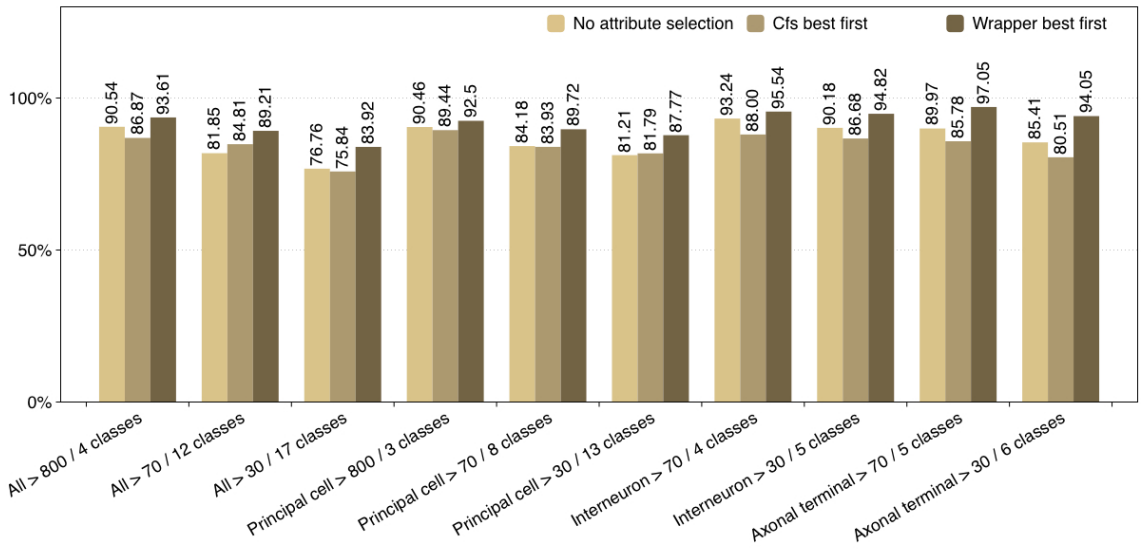


Figura B.3: Clasificación de la especie para el clasificador IB1

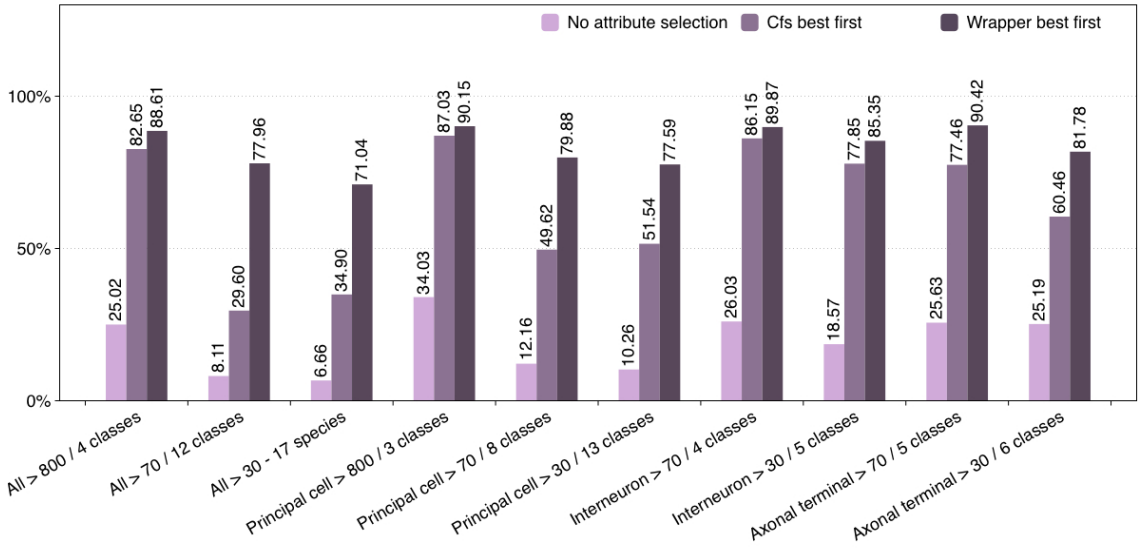


Figura B.4: Clasificación de la especie para el clasificador SVM

B.2. Clasificación del género

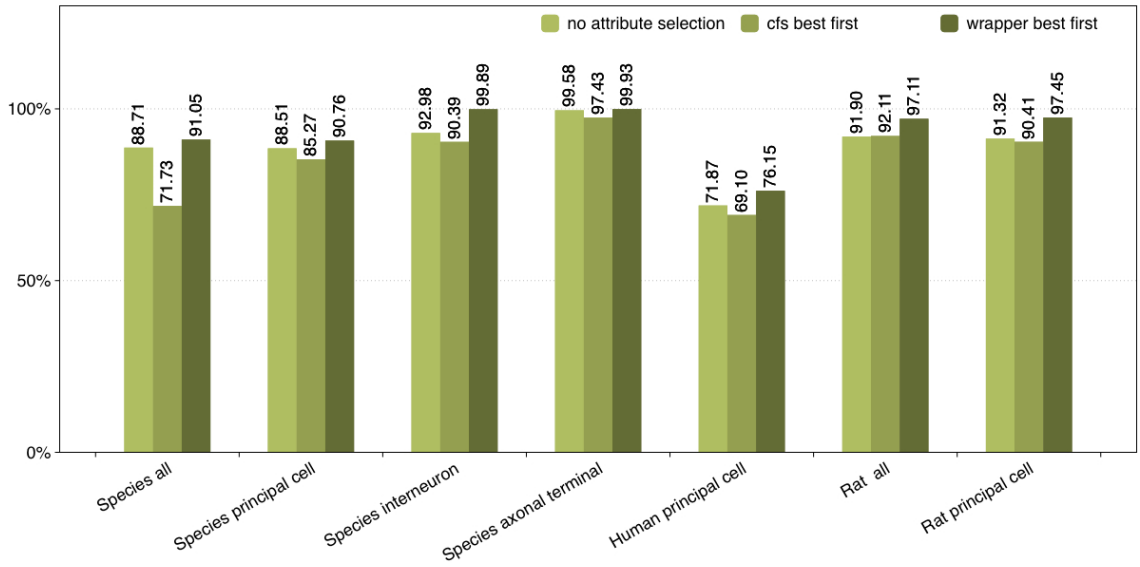


Figura B.5: Clasificación por género para el clasificador bayesiano naïve

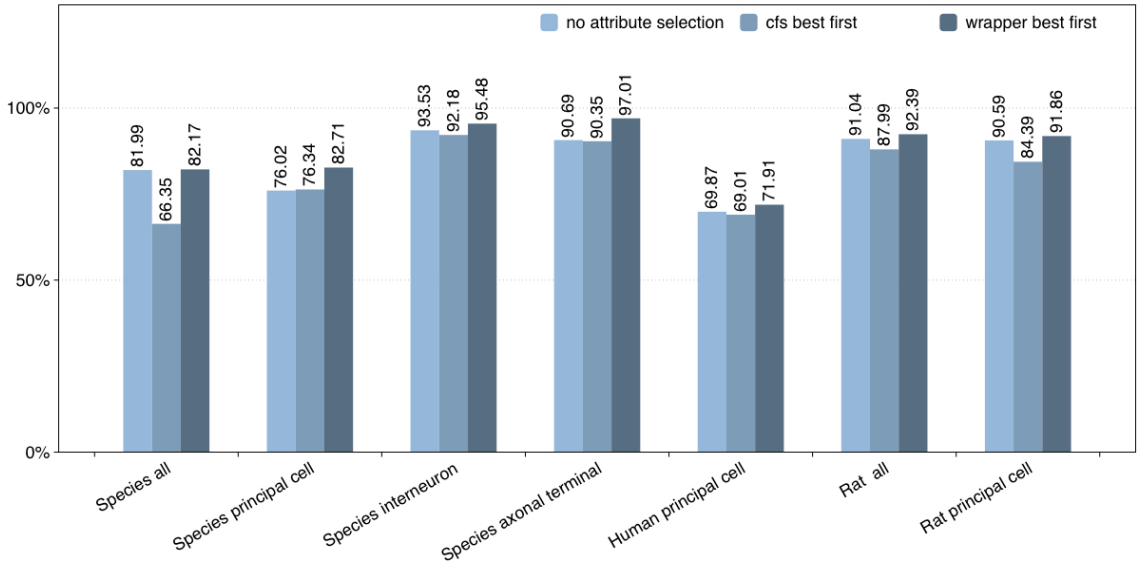


Figura B.6: Clasificación por género para el clasificador j48

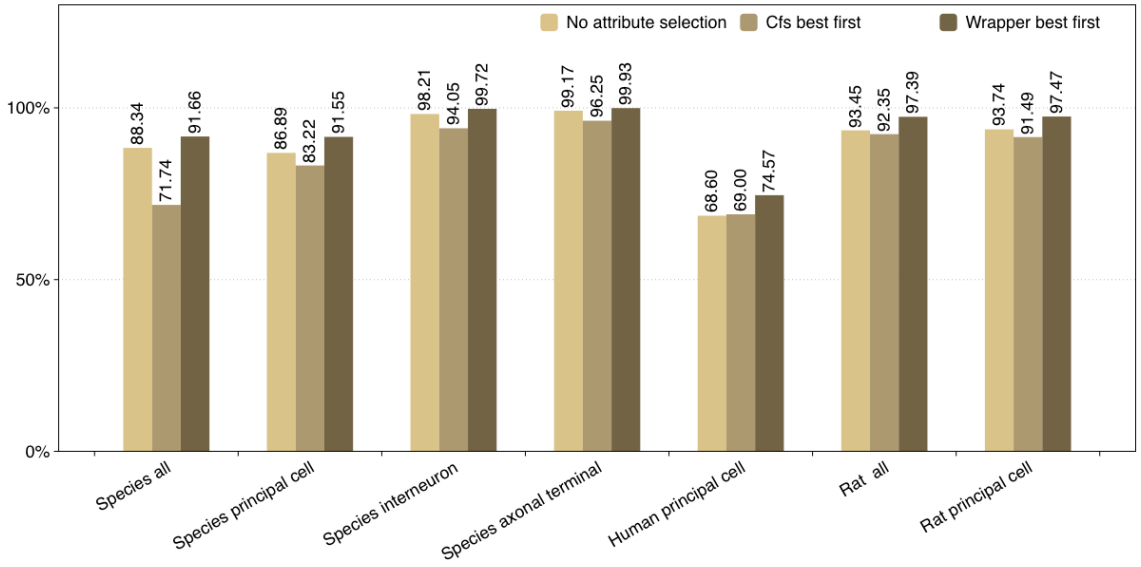


Figura B.7: Clasificación por género para el clasificador IB1

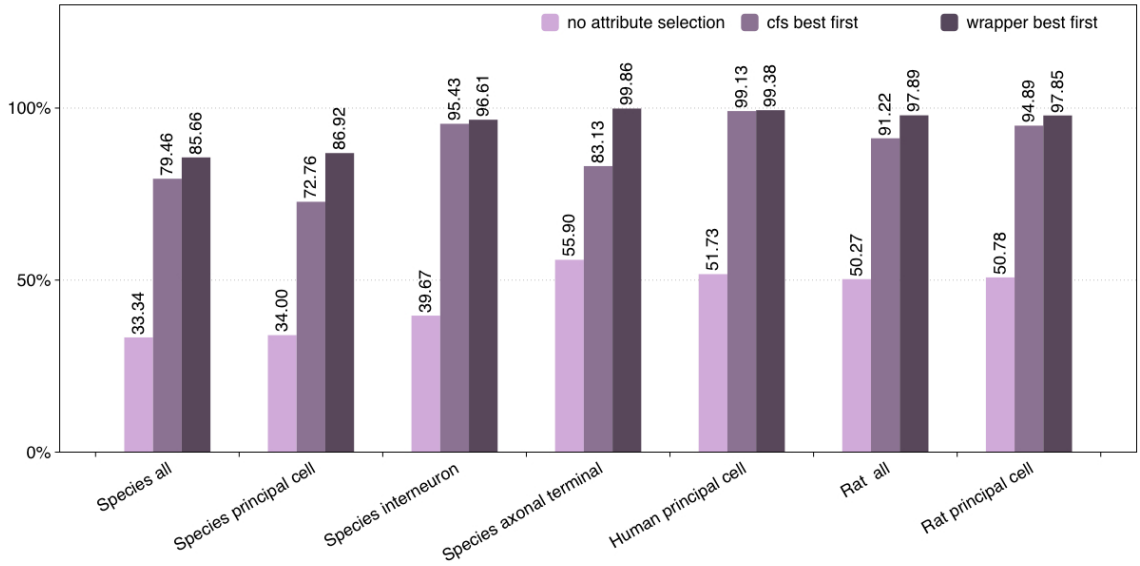


Figura B.8: Clasificación por género para el clasificador SVM

B.3. Clasificación de la edad

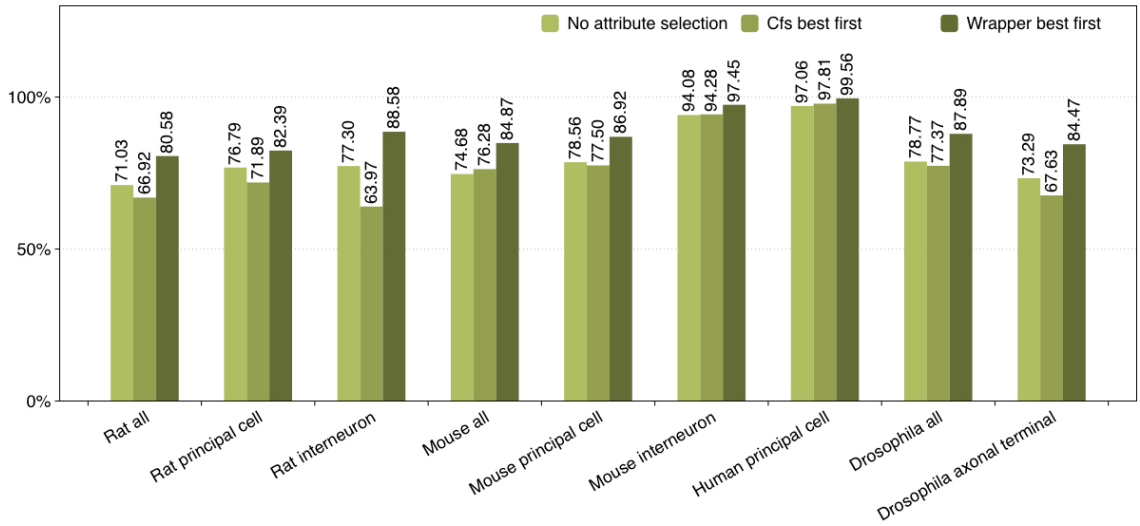


Figura B.9: Clasificación por edad para el clasificador bayesiano naïve

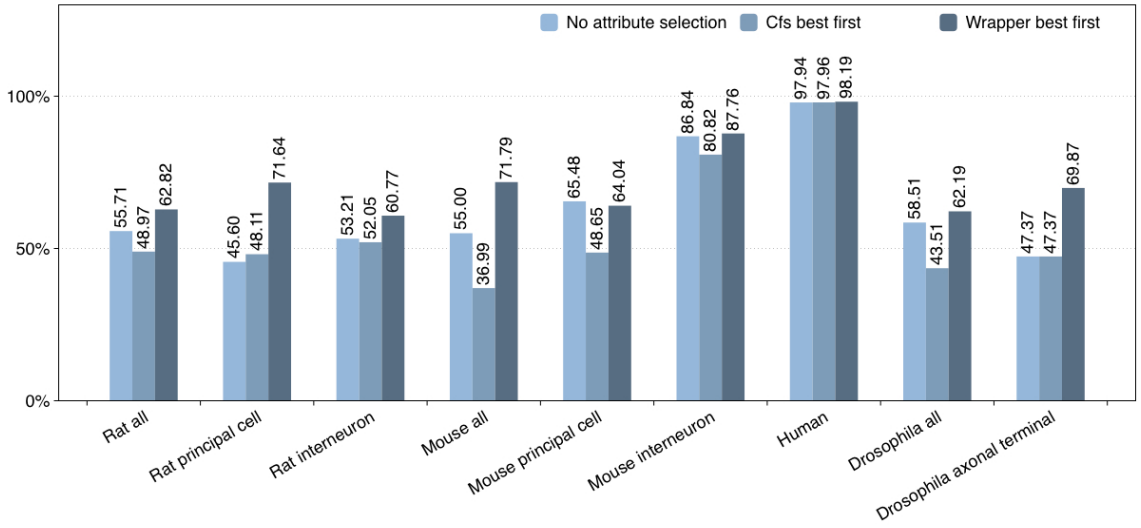


Figura B.10: Clasificación por edad para el clasificador j48

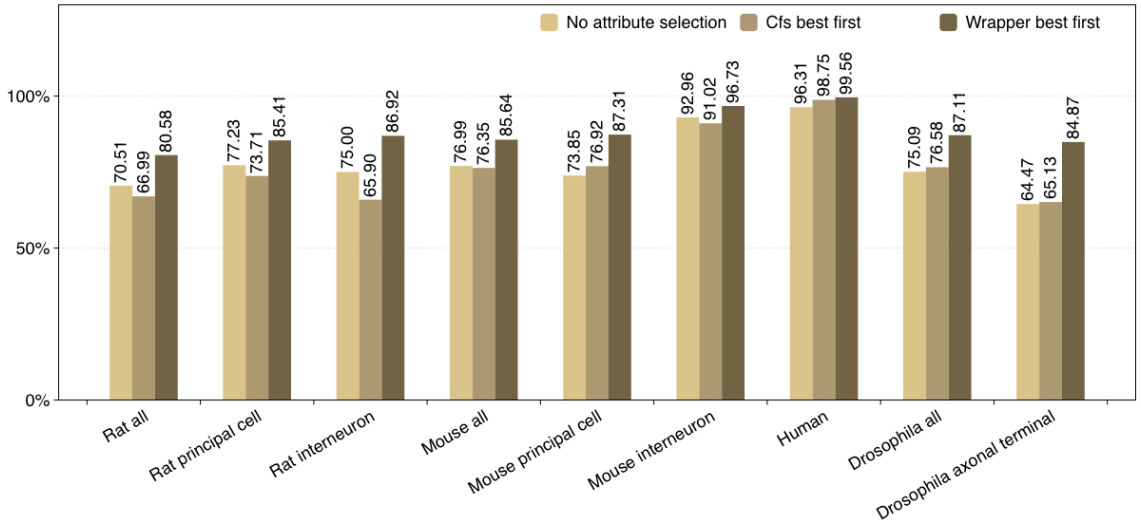


Figura B.11: Clasificación por edad para el clasificador IB1

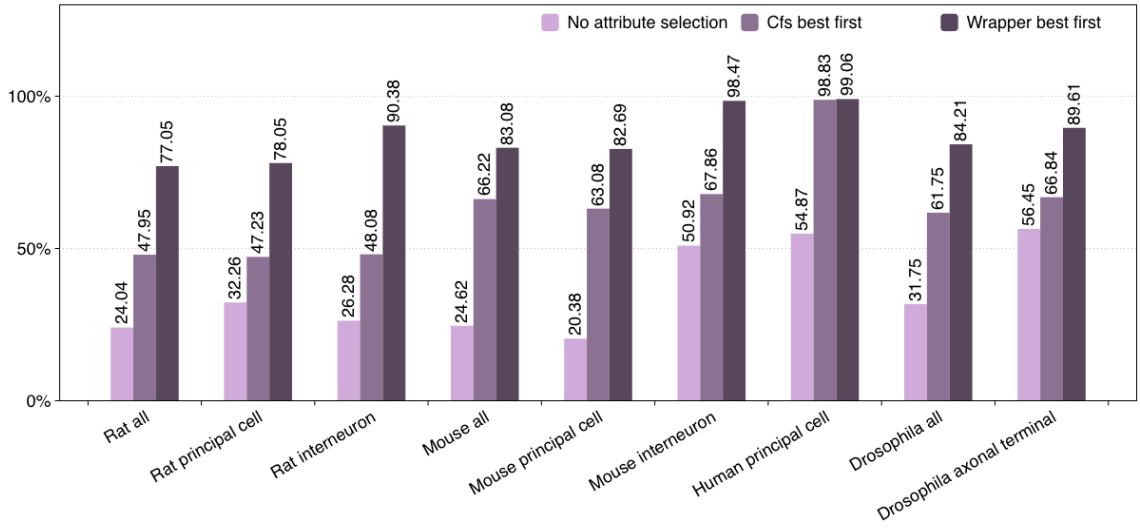


Figura B.12: Clasificación por edad para el clasificador SVM

B.4. Clasificación del tipo de célula

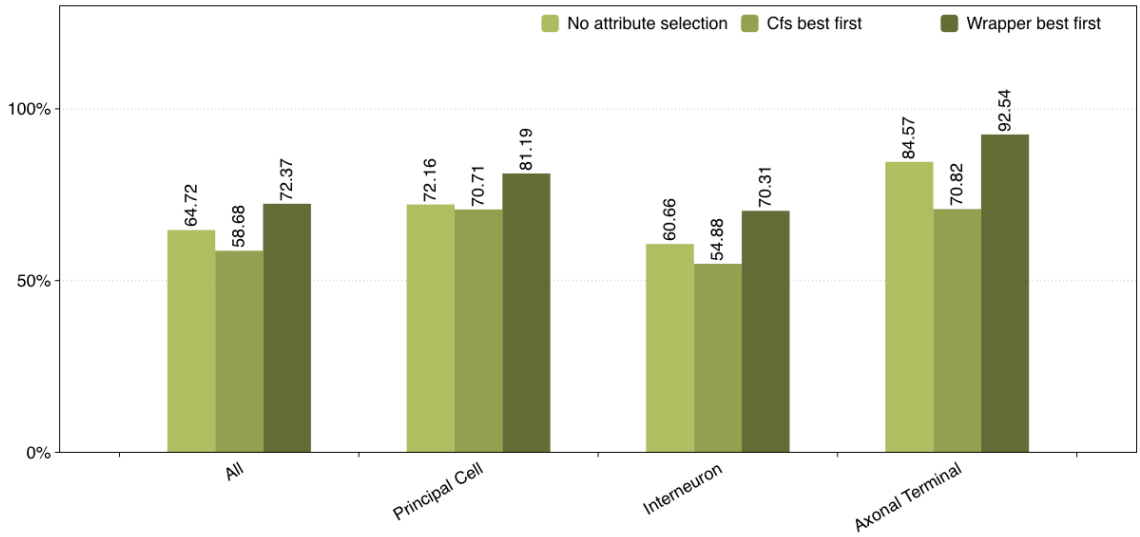


Figura B.13: Clasificación por tipo de célula para el clasificador bayesiano naïve

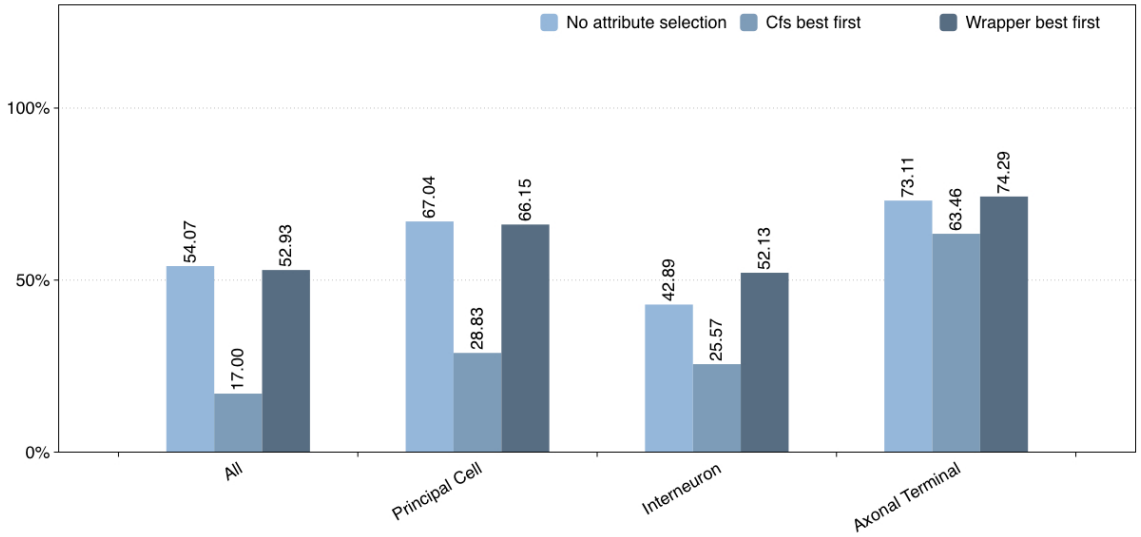


Figura B.14: Clasificación tipo de célula para el clasificador j48

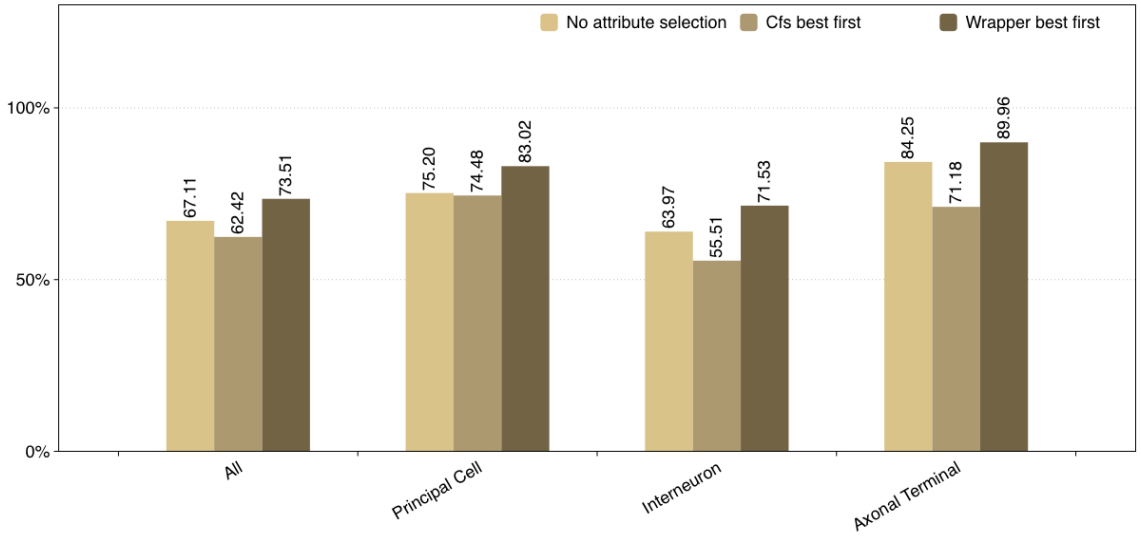


Figura B.15: Clasificación por tipo de célula para el clasificador IB1

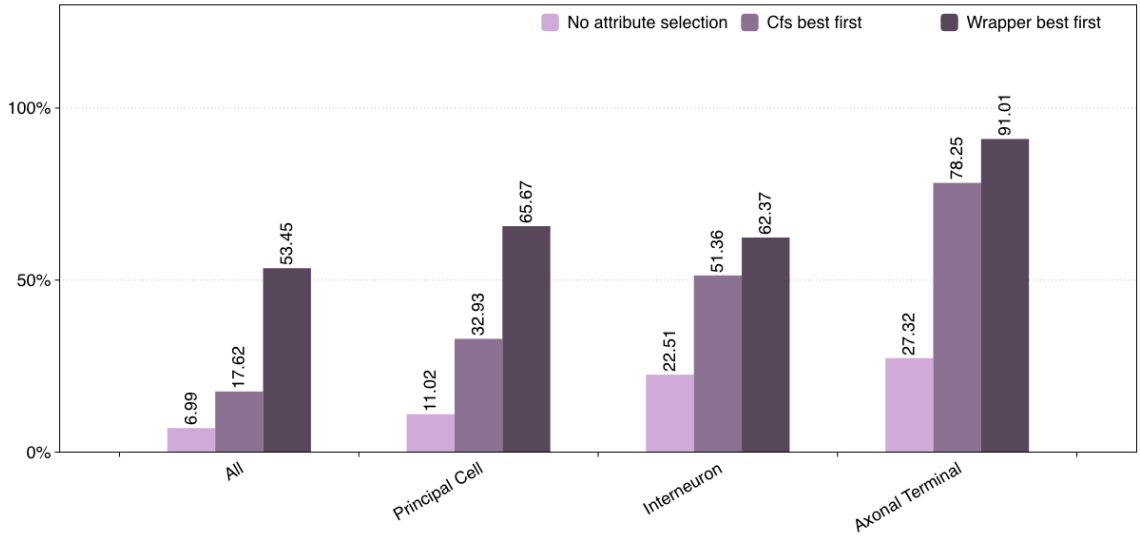


Figura B.16: Clasificación por tipo de célula para el clasificador SVM

B.5. Clasificación por región del cerebro

B.5.1. Clasificación general

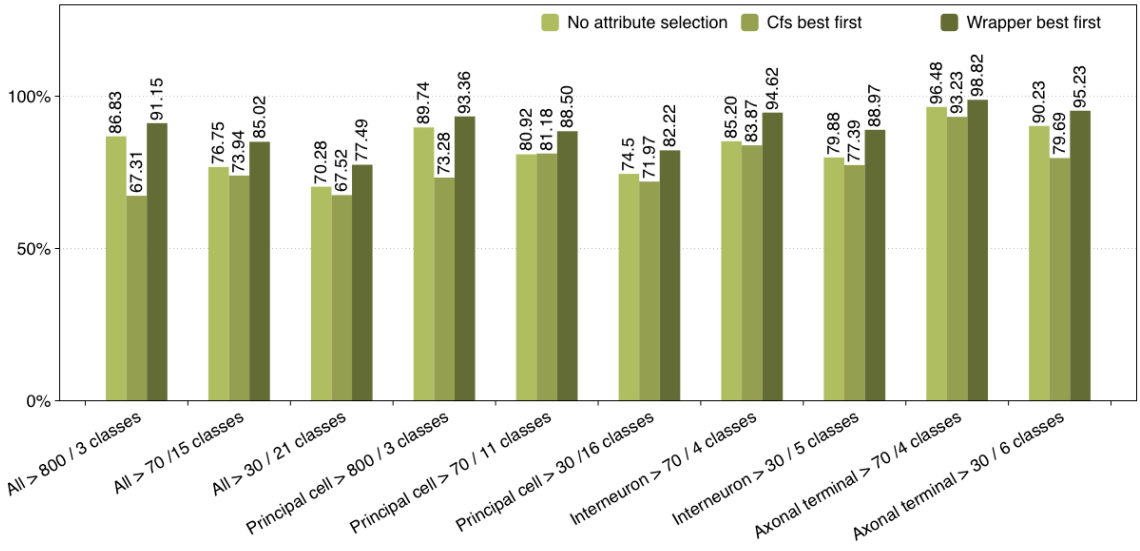


Figura B.17: Clasificación por región del cerebro para el clasificador bayesiano naïve

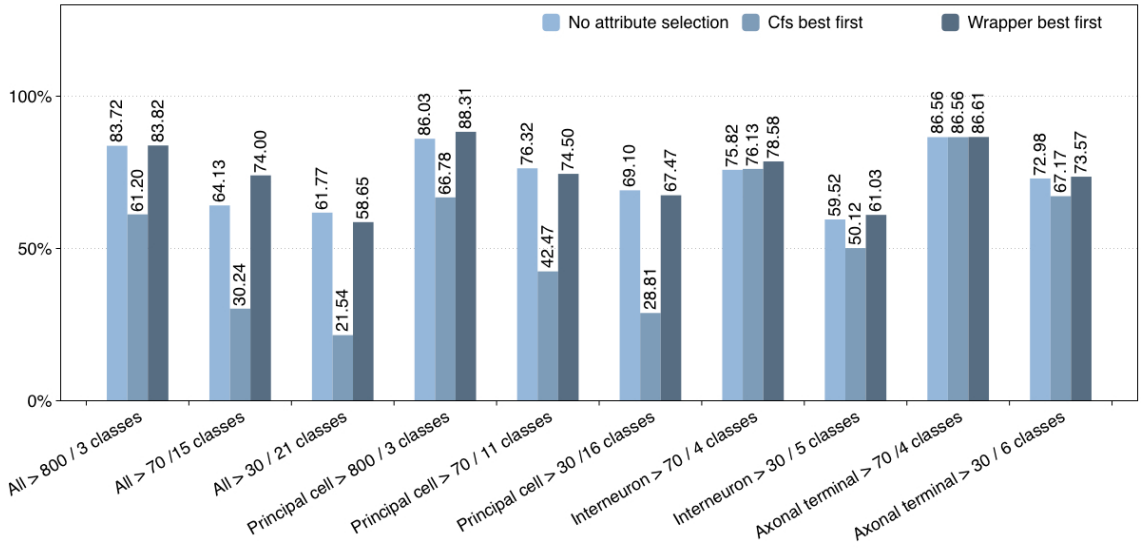


Figura B.18: Clasificación por región del cerebro para el clasificador j48

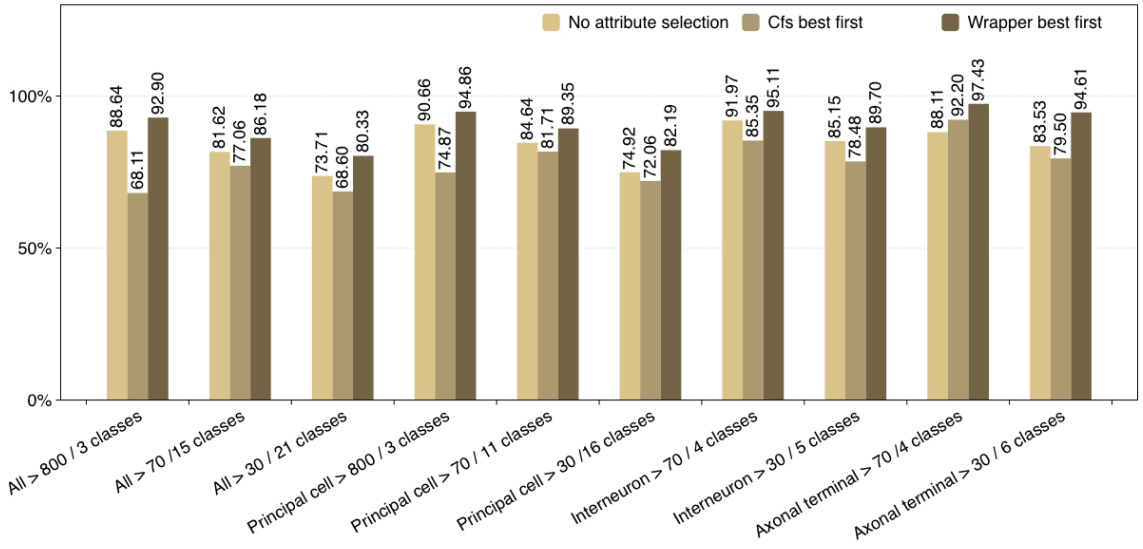


Figura B.19: Clasificación por región del cerebro para el clasificador IB1

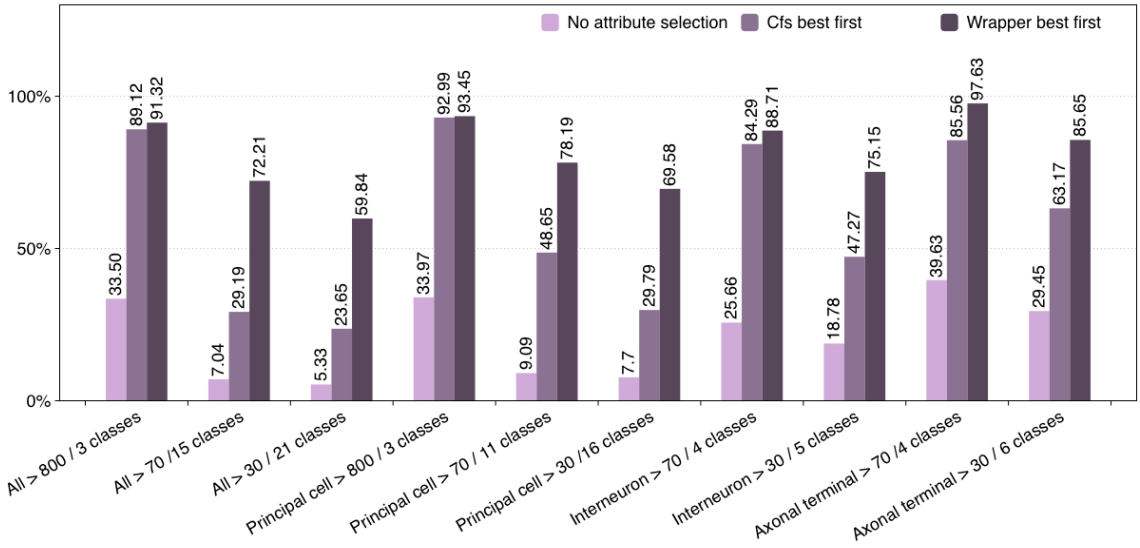


Figura B.20: Clasificación por región del cerebro para el clasificador SVM

B.5.2. Clasificación por región del neocórtex

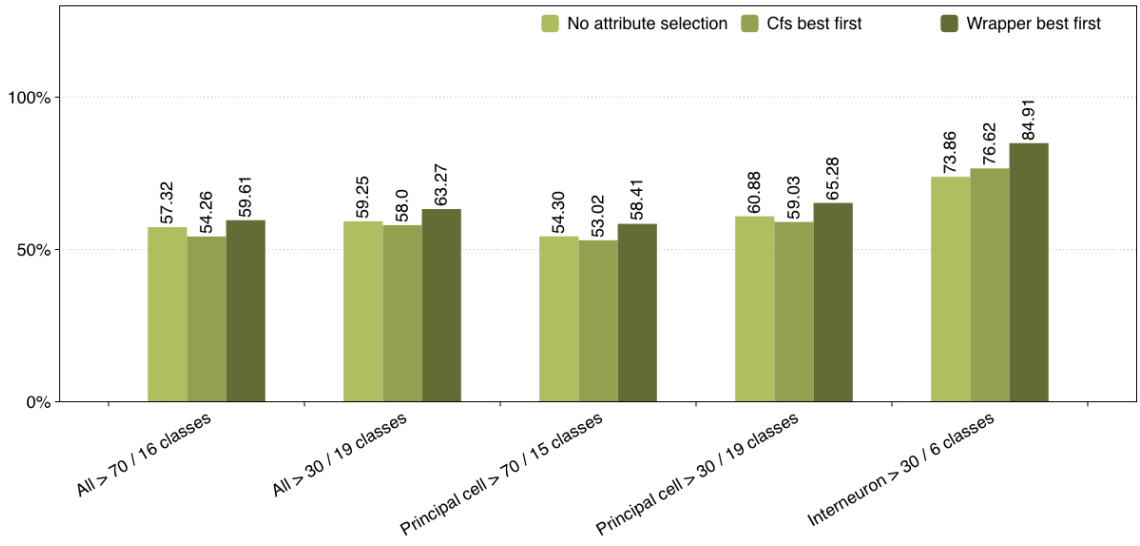


Figura B.21: Clasificación del neocórtex para el clasificador bayesiano naïve

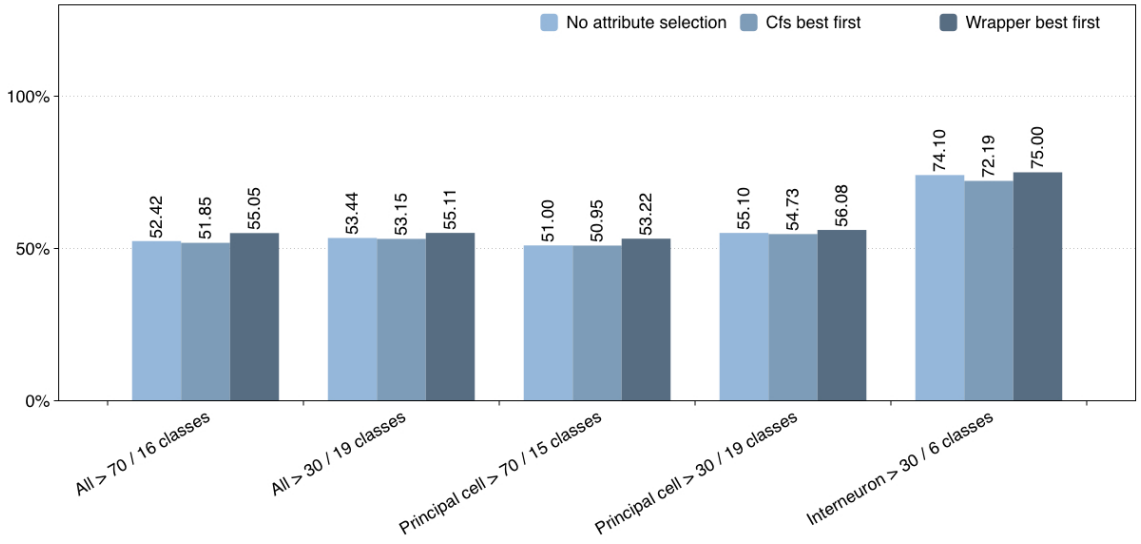


Figura B.22: Clasificación del neocórtex para el clasificador j48

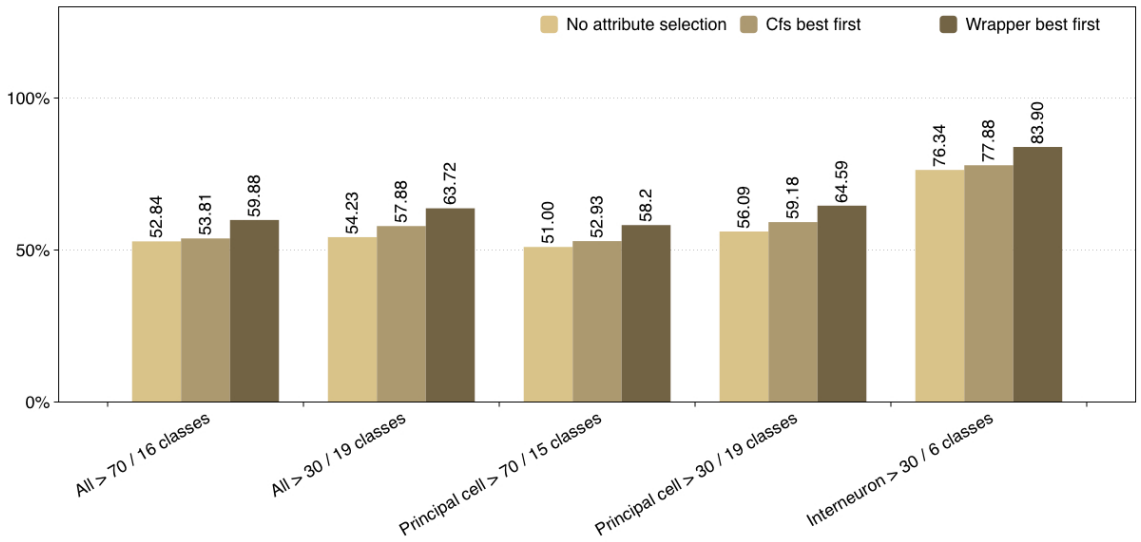


Figura B.23: Clasificación del neocórtex para el clasificador IB1

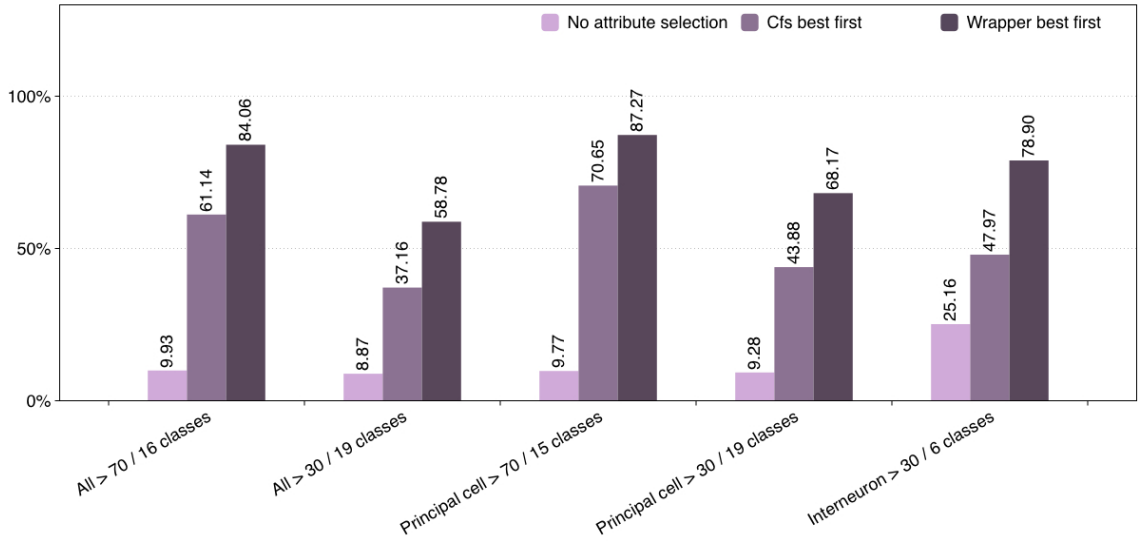


Figura B.24: Clasificación del neocórtex para el clasificador SVM

Anexo C

Bibliografía

Aha, D. W., Kibler, D., y Albert, M. K. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

Ascoli, G. A., Donohue, D. E., y Halavi, M. Neuromorpho. org: a central resource for neuronal morphologies. *The Journal of Neuroscience*, 27(35):9247–9251, 2007.

Ayoub, D. M., Greenough, W. T., y Juraska, J. M. Sex differences in dendritic structure in the preoptic area of the juvenile macaque monkey brain. *Science*, 219(4581):197–198, 1983.

Bielza, C. y Larrañaga, P. Discrete bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)*, 47(1):5, 2014.

Breiman, L. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Burke, S. N. y Barnes, C. A. Neural plasticity in the ageing brain. *Nature Reviews Neuroscience*, 7(1):30–40, 2006.

Buzsáki, G. y Chrobak, J. J. Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Current opinion in neurobiology*, 5(4):504–510, 1995.

Cannon, R., Turner, D., Pyapali, G., y Wheal, H. An on-line archive of reconstructed hippocampal neurons. *Journal of neuroscience methods*, 84(1):49–54, 1998.

Cauli, B., Porter, J. T., Tsuzuki, K., Lambolez, B., Rossier, J., Quenet, B., y Audinat, E. Classification of fusiform neocortical interneurons based on unsupervised clustering. *Proceedings of the National Academy of Sciences*, 97(11):6144–6149, 2000.

Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:341–378, 2002.

Chunwen, L., Xiaqing, X., y Xu, W. A universal neuronal classification and naming scheme based on the neuronal morphology. In *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*, volume 3, pages 2083–2087. IEEE, 2011.

Cortes, C. y Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

De Brabander, J., Kramers, R., y Uylings, H. Layer-specific dendritic regression of pyramidal cells with ageing in the human prefrontal cortex. *European Journal of Neuroscience*, 10(4): 1261–1269, 1998.

DeFelipe, J., López-Cruz, P. L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., Cauli, B., Fairen, A., Feldmeyer, D., *et al.* New insights into the classification and nomenclature of cortical gabaergic interneurons. *Nature Reviews Neuroscience*, 14(3):202–216, 2013.

Druckmann, S., Hill, S., Schurmann, F., Markram, H., y Segev, I. A hierarchical structure of cortical interneuron electrical diversity revealed by automated statistical analysis. *Cerebral Cortex*, 23(12):2994–3006, 2013.

Duda, R. O., Hart, P. E., y Stork, D. G. *Pattern classification*. John Wiley & Sons,, 1999.

Dugger, B. N., Morris, J. A., Jordan, C. L., y Breedlove, S. M. Androgen receptors are required for full masculinization of the ventromedial hypothalamus (vmh) in rats. *Hormones and behavior*, 51(2):195–201, 2007.

Edmunds, S. M. y Parnavelas, J. G. Retzius-cajal cells: an ultrastructural study in the developing visual cortex of the rat. *Journal of neurocytology*, 11(3):427–446, 1982.

Fawcett, T. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Fengqing, H. y Jie, Z. Research for neuron classification based on support vector machine. In *Digital Manufacturing and Automation (ICDMA), 2012 Third International Conference on*, pages 646–649. IEEE, 2012.

Freund, T. F. y Buzsáki, G. Interneurons of the hippocampus. *Hippocampus*, 6(4):347–470, 1996.

Goldstein, L. A., Kurz, E., y Sengelaub, D. R. Androgen regulation of dendritic growth and retraction in the development of a sexually dimorphic spinal nucleus. *The Journal of Neuroscience*, 10(3):935–946, 1990.

Griffin, G. D. y Flanagan-Cato, L. M. Sex differences in the dendritic arbor of hypothalamic ventromedial nucleus neurons. *Physiology & behavior*, 97(2):151–156, 2009.

Grill, J. D. y Riddle, D. R. Age-related and laminar-specific dendritic changes in the medial frontal cortex of the rat. *Brain research*, 937(1):8–21, 2002.

Guerra, L., McGarry, L. M., Robles, V., Bielza, C., Larrañaga, P., y Yuste, R. Comparison between supervised and unsupervised classifications of neuronal cell types: a case study. *Developmental neurobiology*, 71(1):71–82, 2011.

Guerra, L., Benavides-Piccione, R., Bielza, C., Robles, V., DeFelipe, J., y Larrañaga, P. Semi-supervised projected clustering for classifying gabaergic interneurons. In *Artificial Intelligence in Medicine*, pages 156–165. Springer, 2013.

Gulyas, A., Toth, K., McBain, C., y Freund, T. Stratum radiatum giant cells: a type of principal cell in the rat hippocampus. *European Journal of Neuroscience*, 10(12):3813–3822, 1998.

Halavi, M., Polavaram, S., Donohue, D. E., Hamilton, G., Hoyt, J., Smith, K. P., y Ascoli, G. A. Neuromorpho. org implementation of digital neuroscience: dense coverage and integration with the nif. *Neuroinformatics*, 6(3):241–252, 2008.

Hall, M. A. Feature selection for discrete and numeric class machine learning. 1999a.

Hall, M. A. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999b.

Jacobs, B., Johnson, N. L., Wahl, D., Schall, M., Maseko, B. C., Lewandowski, A., Raghanti, M. A., Wicinski, B., Butti, C., Hopkins, W., *et al.* Corrigendum: Comparative neuronal morphology of the cerebellar cortex in afrotherians, carnivores, cetartiodactyls, and primates. *Frontiers in neuroanatomy*, 8:69, 2014.

John, G. H. y Langley, P. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

Juraska, J. M., Fitch, J. M., y Washburne, D. L. The dendritic morphology of pyramidal neurons in the rat hippocampal ca3 area. ii. effects of gender and the environment. *Brain research*, 479(1):115–119, 1989.

Kalsbeek, A., Voorn, P., y Buijs, R. M. Development of dopamine-containing systems in the cns. *Handbook of chemical neuroanatomy*, 10:63–112, 1992.

Karagiannis, A., Gallopin, T., David, C., Battaglia, D., Geoffroy, H., Rossier, J., Hillman, E. M., Staiger, J. F., y Cauli, B. Classification of npy-expressing neocortical interneurons. *The Journal of Neuroscience*, 29(11):3642–3659, 2009.

Kohavi, R. y John, G. H. Wrappers for feature subset selection. *Artificial intelligence*, 97(1): 273–324, 1997.

Konishi, M. y Akutagawa, E. Neuronal growth, atrophy and death in a sexually dimorphic song nucleus in the zebra finch brain. *Nature*, 315:145–147, 1985.

Kubat, M., Matwin, S., *et al.* Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

Kusicka, C. y Schulz, E. Quantitative examinations of stellate cell types of the regio cingularis mesoneocorticalis in comparison to the adjacent regions (mesoarchicortex and neocortex) after deafferentation in the rat. *Journal für Hirnforschung*, 22(6):591, 1981.

Laurikkala, J. *Improving identification of difficult small classes by balancing class distribution*. Springer, 2001.

Marin, E. C., Jefferis, G. S., Komiyama, T., Zhu, H., y Luo, L. Representation of the glomerular olfactory map in the drosophila brain. *Cell*, 109(2):243–255, 2002.

Markham, J., McKian, K., Stroup, T., y Juraska, J. Sexually dimorphic aging of dendritic morphology in ca1 of hippocampus. *Hippocampus*, 15(1):97–103, 2005.

Markham, J. A. y Juraska, J. M. Aging and sex influence the anatomy of the rat anterior cingulate cortex. *Neurobiology of aging*, 23(4):579–588, 2002.

McGarry, L. M., Packer, A. M., Fino, E., Nikolenko, V., Sippy, T., y Yuste, R. Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Frontiers in neural circuits*, 4, 2010.

Murphey, R. y Chiba, A. Assembly of the cricket cercal sensory system: genetic and epigenetic control. *Journal of neurobiology*, 21(1):120–137, 1990.

Nieuwenhuys, R. The neocortex. *Anatomy and embryology*, 190(4):307–337, 1994.

Nieuwenhuys, R., Voogd, J., y Van Huijzen, C. *The human central nervous system: a synopsis and atlas*. Springer, 2007.

Perry, V., Oehler, R., y Cowey, A. Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey. *Neuroscience*, 12(4):1101–1123, 1984.

Possidente, D. y Murphey, R. Genetic control of sexually dimorphic axon morphology indrosophilasensory neurons. *Developmental biology*, 132(2):448–457, 1989.

Purves, D. y Lichtman, J. W. Geometrical differences among homologous neurons in mammals. *Science*, 228(4697):298–302, 1985.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Quinlan, J. R. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

Ramón y Cajal, S. *La rétine des vertébrés*. Typ. de Joseph van In & Cie., 1893.

Saeys, Y., Inza, I., y Larrañaga, P. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

Schuldiner, O. Axon and dendrite pruning in drosophila fengwei yu and oren schuldiner 2. *Current Opinion in Neurobiology*, 27:192–198, 2014.

Schulz, E., Schonheit, B., y Holz, L. Quantitative study of dendritic tree of large (regular) pyramidal cells of lamina v in rat anterior cingular cortex. *Journal für Hirnforschung*, 17 (2):155, 1976.

Scorcioni, R., Polavaram, S., y Ascoli, G. A. L-measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. *Nature protocols*, 3(5):866–876, 2008.

Spruston, N. Pyramidal neurons: dendritic structure and synaptic integration. *Nature Reviews Neuroscience*, 9(3):206–221, 2008.

Tobet, S. y Fox, T. Sex differences in neuronal morphology influenced hormonally throughout life. In *Sexual differentiation*, pages 41–83. Springer, 1992.

Tsiola, A., Hamzei-Sichani, F., Peterlin, Z., y Yuste, R. Quantitative morphologic classification of layer 5 neurons from mouse primary visual cortex. *Journal of Comparative Neurology*, 461(4):415–428, 2003.

Uylings, H. y De Brabander, J. Neuronal changes in normal human aging and alzheimer's disease. *Brain and cognition*, 49(3):268–276, 2002.

Witten, I. H. y Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Wong, A. M., Wang, J. W., y Axel, R. Spatial representation of the glomerular map in the drosophila protocerebrum. *Cell*, 109(2):229–241, 2002.

Xianhua, X. Application of bayes discriminant in neuronal morphology's classification. *Journal of Convergence Information Technology*, 6(8), 2011.

Yu, C., Han, Z., Zeng, W., Liu, S., *et al.* Morphology cluster and prediction of growth of human brain pyramidal neurons. *Neural Regeneration Research*, 7(1):36, 2012.

Zaitsev, A. V. Classification and function of gabaergic interneurons of the mammalian cerebral cortex. *Biochemistry (Moscow) Supplement Series A: Membrane and Cell Biology*, 7(4): 245–259, 2013.

Zawadzki, K., Feenders, C., Viana, M. P., Kaiser, M., y Costa, L. d. F. Morphological homogeneity of neurons: Searching for outlier neuronal cells. *Neuroinformatics*, 10(4):379–389, 2012.