Master Universitario en Bioinformática para Ciencias de la Salud
Facultad de Informática
UNIVERSIDADE DA CORUÑA

Informe de las Prácticas 1 y 2.

INTELIGENCIA COMPUTACIONAL PARA BIOINFORMÁTICA.

Pedro Sánchez García | Andrea Santisteban Veiga
MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA PARA CIENCIAS DE LA SALUD.
PROFESORA: DRA. BEATRIZ PÉREZ SÁNCHEZ.

PRÁCTICA 1. MÉTODOS DE APRENDIZAJE ESTADÍSTICO Y DISEÑO EXPERIMENTAL

CONJUNTO DE DATOS: IRIS

PREPROCESADO DE DATOS: NORMALIZACIÓN.

Se toma como punto de partida el conjunto de datos de Iris disponible en el UCI Machine Learning Repository, planteándose un problema de clasificación que presenta como fundamento la determinación de la pertenencia o no de una entrada a una determinada clase. Esta última se encuentra, junto con otras posibles, predefinidas en un conjunto de observaciones realizadas, de tal forma que el objetivo radica en aprender a asignar un nuevo caso que se presente a una de dichas clases.

En base a la formación del conjunto de datos de Iris, se efectúa un aprendizaje supervisado, con un conjunto de entrenamiento en el que para cada uno de los datos hay una cierta etiqueta, tratando de alcanzar un modelo genérico con parámetros ajustables y minimizar el error comparando la salida de clase con la que se debería obtener (proporcionada durante el entrenamiento). Para ello, se emplea el modelo de discriminante lineal como aproximación estadística para las clasificaciones. Posteriormente, se utilizará el modelo de discriminante cuadrático, donde las medidas de rendimiento y el test estadístico determinarán la decisión del modelo por el que se opta.

Una vez evaluados los modelos de aprendizaje, se prepara el conjunto de datos en la fase de preprocesado de datos del diseño experimental. En el caso del conjunto de Iris, dado que los valores de anchura de sépalo son notablemente reducidos en algunos casos con respecto a los valores del resto de variables (longitud de pétalo, anchura de pétalo y longitud de sépalo), se procede a realizar una normalización. Como consecuencia, los valores del conjunto de datos presentarán unos márgenes determinados, reduciendo un posible sesgo en el modelo de aprendizaje.

APRENDIZAJE.

CONFIGURACIÓN DEL MÉTODO DE APRENDIZAJE.

MODELO DE DISCRIMINANTE LINEAL.

El fundamento de este modelo es generar una división del espacio de entrada mediante hiperplanos que son combinaciones lineales de los atributos de entrada:

$$g = \Sigma_{j=1}^{n} w_j x_j + b, \Theta =< w, b >$$

La ecuación anterior corresponde a la del modelo, donde las x representan las entradas, que se ponderan a través de un valor w, sumando posteriormente un sesgo denotado por b. Por tanto, los parámetros w y b hacen referencia a la pendiente y origen de la recta respectivamente, la cual permite la división del espacio de entrada y clasificación correspondiente.

MODELO DE DISCRIMINANTE CUADRÁTICO.

La esencia de este modelo es similar a la del discriminante lineal, aunque se lleva a cabo por cada clase, es decir, el cálculo del discriminante es la probabilidad o no de pertenencia a una clase que se analiza. A pesar de que se trata de un método sencillo, el principal problema que presenta es el incremento en el número de parámetros a estimar con respecto al discriminante lineal:

$$f_i(x) = \frac{1}{\sqrt{|2\pi_i \sum_i|}} \cdot \exp\left(\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)\right)$$

$$\log \pi_i f_i(x) = \log(\pi_i) - \frac{1}{2}\log\left(\left|\sum_i\right|\right) - \frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)$$

Debemos tener en cuenta que se distingue una función de densidad de probabilidad f_i para la clase C_i, que sigue una distribución normal. De las expresiones anteriores correspondientes a la función de densidad de probabilidad y al discriminante de la clase respectivamente, se alcanzan los parámetros de probabilidad (π_i), media (μ_i) y varianza (x- μ_i). En consecuencia, ante un nuevo dato, se analiza la proyección, que informa del mejor discriminante entre los posibles. Además, cabe destacar que el principal inconveniente de este modelo es que, ante conjuntos reducidos de datos, como en este caso, existe un riesgo elevado de que se produzca un sobreajuste.

MÉTODO DE ESTIMACIÓN DEL ERROR: VALIDACIÓN CRUZADA.

Para llevar a cabo la estimación del error del modelo, dado que el conjunto de Iris es reducido, se opta por la validación cruzada con el fin de aprovechar al máximo el conjunto de datos. El fundamento de este método se basa en la división del conjunto de datos en k subconjuntos disjuntos de tamaño similar. Este parámetro k consta habitualmente de un valor de 10, de forma que en este caso los subconjuntos presentan 135 muestras para entrenamiento y 15 para test. Para cada subconjunto generado, este se selecciona como el conjunto de test para estimar el error real, de tal forma que se entrenan los restantes para el proceso de aprendizaje del modelo. Se calculan los errores sobre el test y se repite el proceso iterando k veces.

MEDIDAS DE RENDIMIENTO.

Tanto para el modelo de discriminante lineal como en el discriminante cuadrático, se han calculado las siguientes medidas de rendimiento: sensibilidad (recall), especificidad, precisión, valor predictivo negativo, exactitud (accuracy) y F1-Score. A continuación, se muestran los resultados correspondientes de cada medida para entrenamiento y test en los modelos, lo que nos permitirá llevar a cabo una rigurosa comparación de estos últimos.

MODELO DE DISCRIMINANTE LINEAL.

Tabla 1. Medidas de rendimiento global para entrenamiento y test del modelo lineal del conjunto de datos Iris.

	Entrenamiento	Test
Sensibilidad (Recall)	0,9800	0,9800
Especificidad	0,9900	0,9900
Precisión	0,9801	0,9833
Valor predictivo negativo	0,9900	0,9909
Exactitud (accuracy)	0,9867	0,9867
F1-Score	0,9800	0,9798

Tal y como se refleja en la Tabla 1, a pesar de las variaciones que se han observado para las medidas de rendimiento de test por clase, se alcanzan unos resultados muy similares en las medidas de rendimiento para entrenamiento y test de forma global. El entrenamiento del discriminante lineal con las entradas correspondientes a cada subconjunto, al igual que el test, muestra en general un buen funcionamiento del modelo, pues todos los valores de las medidas son próximos a 1. No obstante, es preciso destacar que ciertas medidas como la exactitud, que nos proporciona la tasa de acierto global del sistema, no sería adecuada en conjuntos que no se encuentren balanceados. En consecuencia, centramos nuestra atención en la sensibilidad y especificidad, que nos informa de probabilidad de que un caso positivo se clasifique como positivo o que un caso negativo sea clasificado como negativo, respectivamente. En este caso, para ambas medidas, los valores son próximos a 1, de ahí que el modelo entrenado ofrezca un buen comportamiento para la clasificación de Iris.

MODELO DE DISCRIMINANTE CUADRÁTICO.

Tabla 2. Medidas de rendimiento global para entrenamiento y test del modelo cuadrático del conjunto de datos Iris.

	Entrenamiento	Test
Sensibilidad (Recall)	0,9800	0,9733
Especificidad	0,9900	0,9867
Precisión	0,9802	0,9771
Valor predictivo negativo	0,9901	0,9878
Exactitud (accuracy)	0,9867	0,9822
F1-Score	0,9800	0,9728

En lo que respecta al modelo de discriminante cuadrático, se aprecian diferencias de interés en los resultados de las medidas de rendimiento (Tabla 2). En comparación con el modelo de discriminante lineal, se reducen todos los valores alcanzados. Si nos fijamos en las medidas de sensibilidad y especificidad, a pesar de que no resulten cambios muy notables, estas nos informan de que el comportamiento de este modelo entrenado resulta peor para la clasificación de Iris ante nuevos datos proporcionados.

COMPARACIÓN DE MODELOS.

SELECCIÓN DE MEDIDA DE RENDIMIENTO PARA EL TEST ESTADÍSTICO.

Para la realización del test estadístico se han empleado las medidas del F1-Score en lo que respecta al test del modelo de discriminante lineal y cuadrático. Esta medida de rendimiento, en la que nos basamos para comparar los modelos, consiste en la media armónica de la precisión (proporción de casos que se han clasificado correctamente como positivos) y la sensibilidad/recall (probabilidad de que el clasificador nos proporcione un resultado positivo para un caso positivo).

RESULTADOS PARA LOS MODELOS IMPLICADOS.

En la realización del test estadístico, partimos de las medias del F1-Score en test para ambos modelos. A continuación, se muestran las correspondientes desviaciones típicas en cada caso:

Tabla 3. Media y desviación típica de la medida de rendimiento F1-Score para el modelo lineal y cuadrático del conjunto de datos Iris.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	0,9798 ± 0,0592
Discriminante cuadrático	0,9728 ± 0,0592

Se llevó a cabo el test estadístico con un valor crítico de 0.10, que determinará si las muestras correspondientes a las medidas de rendimiento en los modelos resultan o no estadísticamente similares. A continuación, se muestra la tabla ANOVA obtenida del contraste no paramétrico de Kruskal-Wallis:

Tabla 4. Tabla ANOVA de F1-Score del conjunto de datos Iris.

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	5	1	5	0.21	0.6477
Error	450	18	25		
Total	455	19			

Dado que el p-valor obtenido es superior al valor crítico de 0.10 establecido, se mantiene la hipótesis nula de ausencia de diferencias estadísticamente significativas entre los modelos de discriminante lineal y cuadrático.

DIAGRAMA DE CAJAS Y BIGOTES.

En el test estadístico, también se proporciona el resultado correspondiente al diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida elegida para la comparación de modelos:

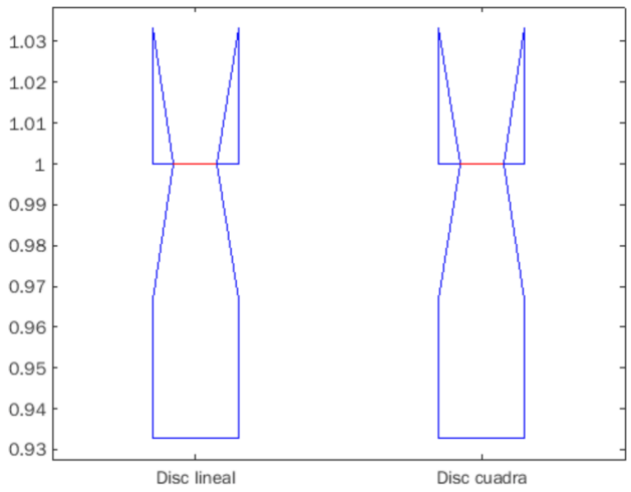


Figura 1. Diagrama de cajas y bigotes para F1-Score del conjunto de datos Iris.

Del diagrama obtenido, debemos tener en cuenta la ausencia de casos atípicos, la posición de primer y tercer cuantil, así como la posición de la mediana en cada modelo para la medida analizada, que en este caso es igual. Por tanto, este resultado muestra y confirma la ausencia de diferencias estadísticamente significativas analizadas con el contraste de Kruskal-Wallis.

DISCUSIÓN Y CONCLUSIONES.

Se han planteado los modelos de discriminante lineal y cuadrático para el conjunto de Iris del UCI Machine Learning Repository. El entrenamiento de los modelos y la posterior estimación del error real de estos mediante validación cruzada conducen a unas medidas de rendimiento que no varían de forma estadísticamente significativa entre ambos modelos. Esto último se ha verificado con el test estadístico a través del contraste de Kruskal-Wallis y el diagrama de cajas y bigotes.

Por otra parte, es preciso indicar que dado el tamaño muestral del conjunto de datos (150 muestras), sería más adecuado el empleo de otra aproximación para la estimación del error de test como el leaving one out, donde se reserva una muestra para test, mientras que las restantes son para el entrenamiento.

En base a los resultados obtenidos y tomando como referencia el principio de la navaja de Occam para minimizar complejidad, se elegiría el modelo de discriminante lineal para la clasificación en Iris.

CONJUNTO DE DATOS: QSAR BIODEGRADATION

PREPROCESADO DE DATOS: NORMALIZACIÓN.

El conjunto de datos QSAR biodegradation disponible en el UCI Machine Learning Repository forma parte de un proyecto financiado por la Comunidad Europea para el desarrollo de modelos QSAR (Quantitative Structure Activity Relationships) en el estudio de relaciones entre la estructura química y la biodegradación de moléculas. Contiene 1055 muestras de compuestos químicos y 41 descriptores moleculares como variables para determinar la clasificación de las entradas a una de las 2 clases: biodegradable / no biodegradable. Al igual que en el conjunto de Iris, se plantea un problema de clasificación. No obstante, en este caso existe un mayor número de variables y se trabaja con 2 clases, con el mismo objetivo de aprender a asignar un nuevo caso que se presente a una de dichas clases con el uso de un modelo.

Teniendo en cuenta el conjunto de datos y el planteamiento del problema, se efectúa un aprendizaje supervisado, distinguiendo un conjunto de entrenamiento para alcanzar el modelo genérico y minimizar el error. Se toma como partida el modelo de discriminante lineal como la aproximación estadística para las clasificaciones. Posteriormente, se utilizará el modelo de discriminante cuadrático y se optará por uno de ellos en función de los resultados para las medidas de rendimiento y del correspondiente test estadístico.

En base a lo anterior, para el diseño experimental del presente caso, la fase de preprocesado de los datos nos conduce a la decisión de efectuar una normalización, ya que, observando el conjunto, se pueden apreciar valores para ciertas variables que son muy distintos en orden de magnitud. Por tanto, se normaliza tratando cada variable de forma independiente con un re escalado de media 0 y desviación típica de 1. Cabe destacar que de este modo se trata la dispersión entre los valores de variables y se reduce un posible sesgo en el modelo de aprendizaje.

APRENDIZAJE.

CONFIGURACIÓN DEL MÉTODO DE APRENDIZAJE.

Tal y como se ha mencionado anteriormente, en cada uno de los modelos se efectúa un proceso de entrenamiento supervisado, proporcionando datos al modelo con el fin de que generalice bien a partir de los datos suministrados y clasifique posteriormente las nuevas entradas. Se trata de un proceso iterativo en el que también se pretenden alcanzar valores óptimos de los parámetros asociados al modelo. Posteriormente, se combina con una estimación del error real del modelo, donde se determina la aproximación a emplear en función del conjunto de datos.

Por tanto, el diseño experimental está basado en 2 conjuntos de datos independientes: entrenamiento y test. El primero se proporciona al modelo para entrenar y ajustar los parámetros asociados, mientras que el segundo será el utilizado para la estimación del error real del test sobre el modelo entrenado, proporcionando datos no empleados en el entrenamiento.

MÉTODO DE ESTIMACIÓN DEL ERROR: VALIDACIÓN CRUZADA.

En base al tamaño muestral del conjunto de datos QSAR, que es superior a 1000 muestras, se ha elegido la validación cruzada como el método para estimar el error. Se divide el conjunto de datos en k=5 subconjuntos disjuntos con un tamaño similar. En este caso, los subconjuntos presentan 844 muestras para entrenamiento y 211 para test. Progresivamente, se seleccionan 4 subconjuntos para el entrenamiento o proceso de aprendizaje y uno de los subconjuntos para el test, mediante el cual se estima el error real del modelo. Se trata de un proceso iterativo, constando de k=5 iteraciones para el presente caso.

MEDIDAS DE RENDIMIENTO.

Al igual que en el conjunto de Iris, tanto para el modelo de discriminante lineal como en el discriminante cuadrático, se han calculado las siguientes medidas de rendimiento: sensibilidad (recall), especificidad, precisión, valor predictivo negativo, exactitud (accuracy) y F1-Score. A continuación, se muestran los resultados de cada medida para entrenamiento y test en los modelos, lo que nos permitirá realizar una comparación de estos últimos antes de proceder al test estadístico:

MODELO DE DISCRIMINANTE LINEAL.

Tabla 5. Medidas de rendimiento global para entrenamiento y test del modelo lineal del conjunto de datos Qsar.

	Entrenamiento	Test
Sensibilidad (Recall)	0,8530	0,8349
Especificidad	0,8530	0,8349
Precisión	0,8631	0,8422
Valor predictivo negativo	0,8631	0,8422
Exactitud (accuracy)	0,8744	0,8569
F1-Score	0,8576	0,8374

Los resultados de las medidas de rendimiento para entrenamiento y test de forma global son similares (Tabla 5). No obstante, se puede observar que las medidas en el caso del entrenamiento no presentan valores próximos a 1, lo que parece indicar que no se produce una generalización adecuada y en consecuencia, un entrenamiento adecuado del modelo. Esto conduce a unos valores en el caso de las medidas para el caso de test que también se distancian de 1.

Al igual que se hizo en el caso del conjunto de Iris, centramos nuestra atención en la sensibilidad y especificidad. En base a estas, no podemos afirmar que el modelo entrenado proporcione un buen comportamiento para la clasificación de los compuestos químicos.

MODELO DE DISCRIMINANTE CUADRÁTICO.

Tabla 6. Medidas de rendimiento global para entrenamiento y test del modelo cuadrático del conjunto de datos Qsar.

	Entrenamiento	Test
Sensibilidad (Recall)	0,8400	0,8060
Especificidad	0,8400	0,8060
Precisión	0,8045	0,7756
Valor predictivo negativo	0,8045	0,7756
Exactitud (accuracy)	0,8097	0,7858
F1-Score	0,8034	0,7768

Para el modelo de discriminante cuadrático, se aprecian notables diferencias entre los resultados de las medidas de rendimiento para entrenamiento y test (Tabla 6). En comparación con el modelo de discriminante lineal, se reducen más los valores alcanzados de las medidas en el caso de test. Atendiendo a las medidas de sensibilidad y especificidad en test, estas nos informan de un peor comportamiento del modelo entrenado para la clasificación de compuestos ante nuevos datos de entrada suministrados.

COMPARACIÓN DE MODELOS.

SELECCIÓN DE MEDIDA DE RENDIMIENTO PARA EL TEST ESTADÍSTICO.

El test estadístico se ha realizado basándose en las medidas del F1-Score en lo que respecta al test de ambos modelos.

RESULTADOS PARA LOS MODELOS IMPLICADOS.

Se muestran las medias y correspondientes desviaciones típicas para las medidas del F1-Score empleadas en el test estadístico para la comparación de los modelos:

Tabla 7. Media y desviación típica de la medida de rendimiento F1-Score para el modelo lineal y cuadrático del conjunto de datos Qsar.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	**0,8374 ± 0,0770**
Discriminante cuadrático	**0,7768 ± 0,0567**

Para el test estadístico se estableció un valor crítico de 0.10, que determina si las muestras correspondientes a las medidas de rendimiento en los modelos resultan o no similares estadísticamente. A continuación, se muestra la correspondiente tabla resultante del ANOVA de 1 vía:

Tabla 8. Tabla ANOVA de F1-Score del conjunto de datos Qsar.

Source	SS	df	MS	F	Prob>F
Columns	0.00661	1	0.00661	17.06	0.0033
Error	0.0031	8	0.00039		
Total	0.00971	9			

El p-valor resultante es superior al valor crítico de 0.10 establecido, de tal manera que se rechaza la hipótesis nula de ausencia de diferencias estadísticamente significativas entre los modelos de discriminante lineal y cuadrático.

DIAGRAMA DE CAJAS Y BIGOTES.

La existencia de diferencias estadísticamente significativas se verifica posteriormente con la representación del diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida para la comparación de modelos:

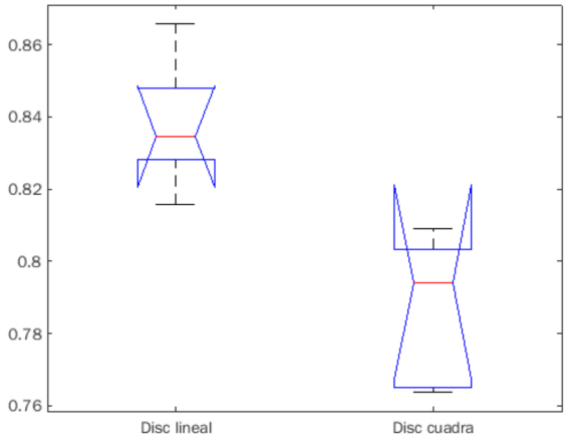


Figura 2. Diagrama de cajas y bigotes para F1-Score del conjunto de datos Qsar.

Del diagrama, se distingue principalmente que el modelo de discriminante lineal presenta un menor rango intercuartílico y una mediana más elevada para la medida de rendimiento analizada en comparación con el discriminante cuadrático.

TEST DE COMPARACIÓN MÚLTIPLE.

La existencia de diferencias estadísticamente significativas entre los dos modelos nos conduce a la realización de un test de comparación múltiple, como una representación complementaria al diagrama de cajas y bigotes:

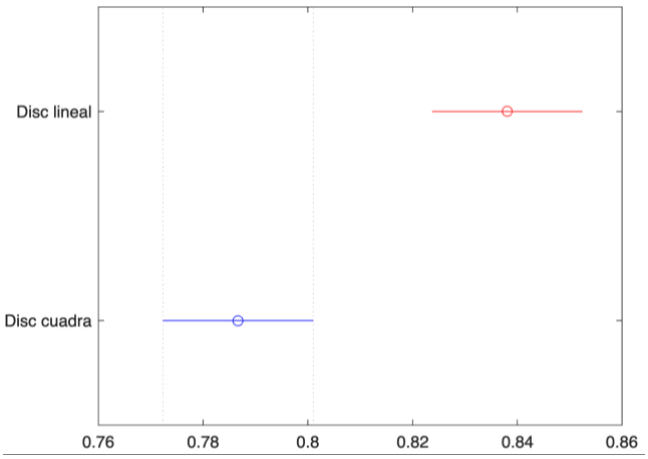


Figura 3. Representación del test de comparación múltiple para el conjunto de datos Iris.

Este nos muestra la tendencia observada anteriormente, de tal forma que si tuviésemos un caso con más de 2 modelos o en el que no nos resultasen claros los grupos entre los que se producen las diferencias, este test de comparación múltiple nos mostraría una descripción para obtener la respuesta.

DISCUSIÓN Y CONCLUSIONES.

El conjunto de datos QSAR del UCI Machine Learning Repository pone de manifiesto unos resultados interesantes. Las medidas de rendimiento en los modelos, tras el entrenamiento y estimación del error real de estos, varían de forma estadísticamente significativa.

El test estadístico correspondiente al ANOVA de 1 vía, el diagrama de cajas y bigotes, así como el test de comparación múltiple nos conduce a esa observación. Sin embargo, la media de la F1-Score analizada es más elevada para el modelo de discriminante lineal. Teniendo en cuenta este aspecto, junto con la "razón de Occam", en un principio se elegiría el modelo de discriminante lineal para la clasificación de nuevos compuestos como datos de entrada.

Además, cabe destacar que, para este conjunto de datos, la reducción de la dimensionalidad se debería incluir como una fase en el preprocesado de los datos. Debemos tener en cuenta que, entre las 41 variables, es posible que se produzca redundancia entre los datos, lo que determinaría la extracción de variables para eliminar el posible ruido en el sistema de aprendizaje del modelo.

PRÁCTICA 2. MÉTODOS DE APRENDIZAJE ESTADÍSTICO, ÁRBOLES DE DECISIÓN Y DISEÑO EXPERIMENTAL

CONJUNTO DE DATOS: IRIS

PREPROCESADO DE DATOS: NORMALIZACIÓN.

Se toma como punto de partida el conjunto de datos de Iris normalizado en la práctica 1. Para esta práctica, nos encontramos en la aproximación de la IA simbólica, donde el conocimiento adquirido como resultado de la simulación de comportamiento inteligente ha de ser comprensible. En este contexto, con respecto a la clasificación del sistema, mediante el ajuste mental, se puede interpretar y formalizar la partición del espacio de entrada en estructuras más simples durante el entrenamiento del modelo. De esta forma, se produce un conocimiento explícito en base al entrenamiento (estructuras lineales y clases asociadas) que nos resulta comprensible.

Al igual que con el modelo de discriminante lineal y cuadrático tratados en la práctica 1, el objetivo es lograr descripciones de las clases para poder clasificar datos nuevos posteriormente. Resulta fundamental determinar aquella partición que explique la totalidad de ejemplos y minimice la complejidad.

APRENDIZAJE.

CONFIGURACIÓN DEL MÉTODO DE APRENDIZAJE.

Se efectúa un proceso de clasificación de Iris mediante árboles de decisión, donde se representa el conocimiento adquirido (estados) como resultado de un algoritmo (CART) basado en el principio de *divide y vencerás*.

El objetivo es lograr 3 versiones de árboles de decisión, seleccionando posteriormente la mejor versión tras valorar las medidas de rendimiento y el resultado gráfico correspondiente. Para ello, es preciso tener en cuenta la estructura de los árboles de decisión. En ellos, el punto de partida está conformado por un conjunto de ejemplos de diferentes clases, del que parten nodos que corresponden con test sobre 1 único atributo y que dividen el conjunto E en subconjuntos E1, E2,…,En. Además, se denotan los nodos hoja como el conjunto de clases disjuntas C1, C2,…, Cn. La selección será en base a las ramas y reglas de la estructura. Finalmente, se aplica el test estadístico, donde se compara con los modelos de discriminante lineal y cuadrático.

Teniendo en cuenta lo anterior, se valora el ajuste del número máximo de divisiones, número mínimo de observaciones en nodo hoja y el número mínimo de ejemplos en 1 rama. Se determina generar las siguientes versiones de los árboles de decisión, con los valores recogidos en la siguiente tabla:

Tabla 9. Valores utilizados para obtener los tres arboles de decisión para el conjunto de datos Iris: número máximo de divisiones, número mínimo de observaciones en nodo hoja y el número mínimo de ejemplos en 1 rama.

Árbol de decisión	Número máximo de divisiones	Número mínimo de observaciones en nodo hoja	Número mínimo de ejemplos en 1 rama
1	149	1	10
2	149	5	15
3	149	10	20

En cada una de las versiones del árbol, con el algoritmo CART (variante del algoritmo ID3), el espacio de estados d-dimensional se subdivide en espacios más reducidos aplicando un discriminante lineal. A partir de un 1 atributo de entrada se establecen tests y si estos superan un criterio de calidad, se establecen en el árbol como nodos hoja para la clasificación.

MÉTODO DE ESTIMACIÓN DEL ERROR: VALIDACIÓN CRUZADA.

Tal y como se recoge anteriormente en el informe de la práctica 1, con el fin de realizar la estimación del error del modelo se determina la validación cruzada con k=10 subconjuntos disjuntos. Mediante k= 10 iteraciones, se destina un subconjunto como el de test para la estimación del error real del modelo y los restantes para el entrenamiento/aprendizaje del modelo.

MEDIDAS DE RENDIMIENTO.

Se han calculado las mismas medidas de rendimiento que en los casos de discriminante lineal y cuadrático tratados en la práctica 1. A continuación, se muestran los resultados correspondientes, que nos servirán, junto con el árbol obtenido, a decidir la mejor versión con la que llevar a cabo la posterior comparación con los modelos de discriminante lineal y cuadrático:

Tabla 10. Medidas de rendimiento para entrenamiento y test de cada árbol obtenido para el conjunto de datos Iris.

	Entrenamiento			Test		
	1	2	3	1	2	3
Sensibilidad (Recall)	**0,9800**	**0,9733**	**0,9607**	**0,9533**	**0,9667**	**0,9467**
Especificidad	**0,9900**	**0,9867**	**0,9804**	**0,9767**	**0,9833**	**0,9733**
Precisión	**0,9810**	**0,9740**	**0,9629**	**0,9589**	**0,9738**	**0,9511**
Valor predictivo negativo	**0,9903**	**0,9868**	**0,9809**	**0,9782**	**0,9854**	**0,9745**
Exactitud (accuracy)	**0,9867**	**0,9822**	**0,9738**	**0,9689**	**0,9778**	**0,9644**
F1-Score	**0,9800**	**0,9733**	**0,9607**	**0,9530**	**0,9659**	**0,9464**

Debido a las características del conjunto de datos, no existen grandes diferencias en las medidas de rendimiento cuando se varía el número de observaciones mínimas en los nodos hoja y en las ramas.

Los resultados anteriores son un aspecto en el que nos centramos para elegir la versión de árbol de cara a la comparación con el modelo de discriminante lineal y cuadrático. Además, estos se complementan con los resultados gráficos alcanzados para cada versión de árbol, mostrados a continuación:

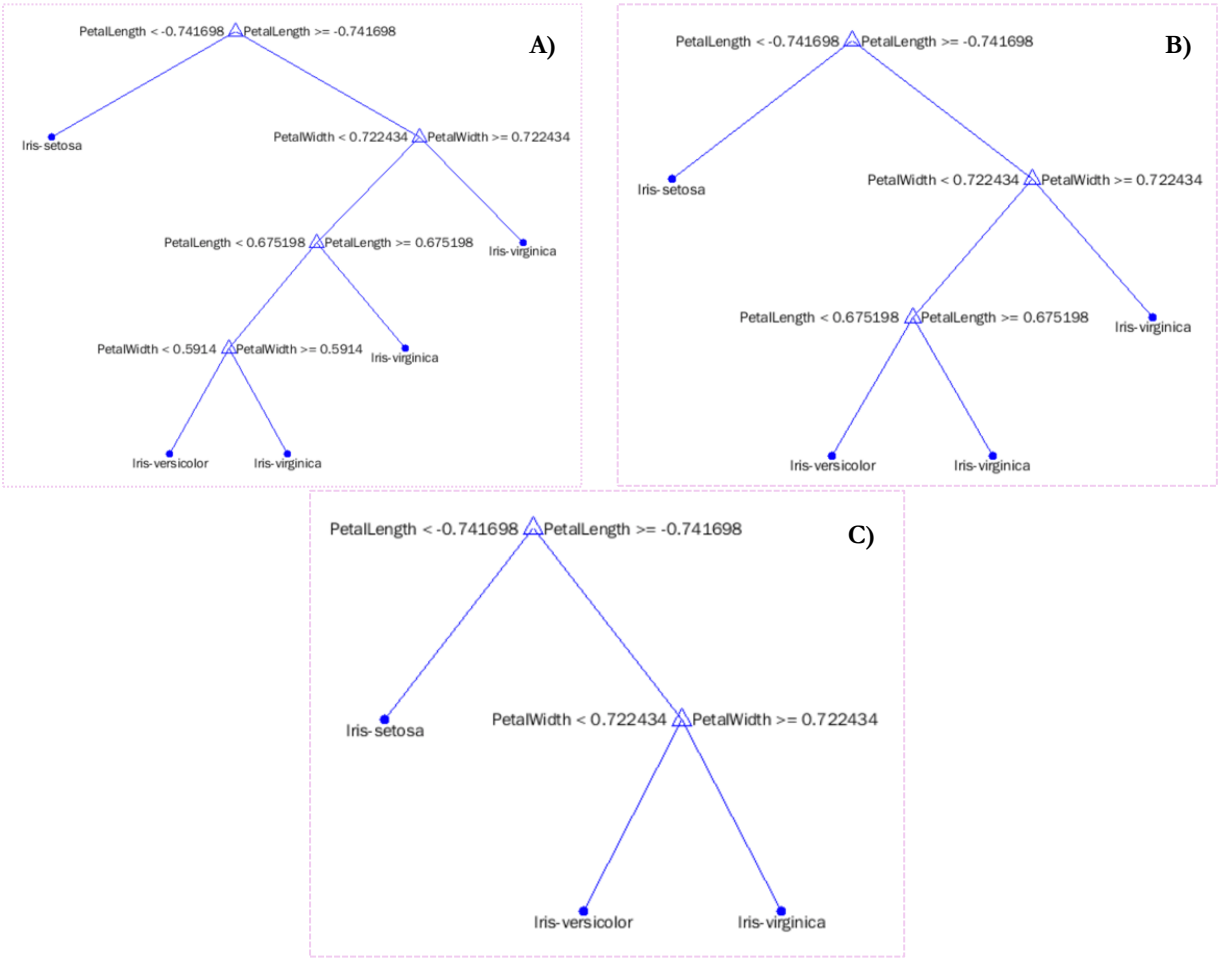


Figura 4. Representación de los tres árboles de decisión obtenidos para el conjunto de datos Iris. **A)** Árbol con una única observación en el nodo hoja y 10 en ramas (valores por defecto). **B)** Árbol con 5 observaciones en nodo hoja y 15 en ramas. **C)** Árbol con 10 observaciones en nodo hoja y 20 en ramas.

Al comparar la versión por defecto (Figura 4.A) frente al caso de 5 observaciones mínimas en nodo hoja y 15 en ramas (Figura4.B), se observa que esta última presenta un aspecto más simple, con un menor número de ramas y una menor redundancia en lo que a reglas se refiere.

Estas características también están presentes en la última versión de árbol obtenida (Figura 4.C), por lo que teniendo en cuenta la representación más simple y las medidas de rendimiento poco variables con respecto a los otros casos, optamos por elegir este árbol para la comparación de modelos.

COMPARACIÓN DE MODELOS.

SELECCIÓN DE MEDIDA DE RENDIMIENTO PARA EL TEST ESTADÍSTICO.

En base a las medidas del F1-Score, se realiza el test estadístico para la comparación de la versión de árbol elegida con los otros modelos.

RESULTADOS PARA LOS MODELOS IMPLICADOS.

Se muestran las medias y correspondientes desviaciones típicas para las medidas del F1-Score empleadas en el test estadístico para la comparación de los modelos:

Tabla 11. Media y desviación típica de la medida de rendimiento F1-Score para el árbol de decisión seleccionado del conjunto de datos Iris.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	0,9798 ± 0,0592
Discriminante cuadrático	0,9728 ± 0,0592
Árbol de decisión	0,9464 ± 0,0527

Con un valor crítico de 0.10, se analiza si las muestras de las medidas de rendimiento en los diferentes modelos resultan o no similares estadísticamente. A continuación, se muestra la tabla ANOVA obtenida del contraste no paramétrico de Kruskal-Wallis:

Tabla 12. Tabla ANOVA de F1-Score del conjunto de datos Iris.

Source	SS	df	MS	F	Prob>F
Columns	163.4	2	81.7	2.74	0.254
Error	1565.6	27	57.9852		
Total	1729	29			

Teniendo en cuenta que el p-valor alcanzado es superior al valor crítico de 0.10 establecido, se rechaza la hipótesis nula de ausencia de diferencias estadísticamente significativas entre los modelos de discriminante lineal, cuadrático y árbol de decisión.

DIAGRAMA DE CAJAS Y BIGOTES.

La ausencia de diferencias estadísticamente significativas determinada se verifica posteriormente con la representación del diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida para la comparación de modelos:

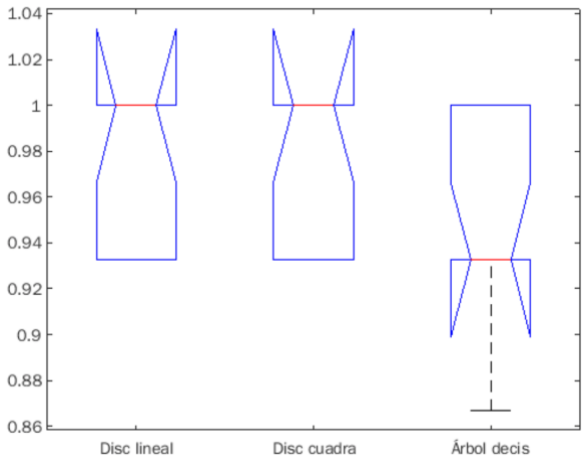


Figura 5. Diagrama de cajas y bigotes para F1-Score del conjunto de datos Iris.

Del diagrama, se aprecia en general que, para el árbol de decisión, hay un mayor rango intercuartílico y una mediana más reducida para la medida de rendimiento analizada en comparación con los otros modelos.

DISCUSIÓN Y CONCLUSIONES.

Se han planteado varias versiones del árbol de decisión para el conjunto de Iris del UCI Machine Learning Repository. En base a las medidas de rendimiento para el entrenamiento y test, junto con el gráfico obtenido, se ha elegido la versión con 10 observaciones mínimas en nodo hoja y 20 en ramas. No obstante, la comparación de los modelos empleando las medidas de F1-Score para test indica ausencia de diferencias estadísticamente significativas entre ellos.

De esta forma, dado que las medidas de rendimiento para el árbol se reducen en comparación con los otros, se valoraría el uso del discriminante lineal para proceso de clasificación en Iris. No obstante, cabe destacar que un mayor número de versiones nos permitiría alcanzar un resultado más comprensible y simplificado, como resultado de una mayor ganancia de información.

CONJUNTO DE DATOS: QSAR BIODEGRADATION

PREPROCESADO DE DATOS: NORMALIZACIÓN.

Se parte del conjunto de datos de QSAR normalizado en la práctica 1. De nuevo, en la aproximación de la IA simbólica, perseguimos aquella partición del espacio de entrada en estructuras simples comprensibles para formalizar la clasificación de nuevos datos en base al entrenamiento efectuado.

APRENDIZAJE.

CONFIGURACIÓN DEL MÉTODO DE APRENDIZAJE.

Con el proceso de clasificación de Iris mediante árboles de decisión, se representa el conocimiento adquirido (estados) como resultado del algoritmo (CART). Se generan 3 versiones de árboles de decisión, seleccionando posteriormente la mejor en base a las medidas de rendimiento y la representación gráfica correspondiente. Para ello, es preciso valorar las ramas y reglas presentes en la estructura de los árboles. Posteriormente, se aplica el test estadístico para comparar con los modelos de discriminante lineal y cuadrático.

Teniendo en cuenta lo anterior, se valora el ajuste del número máximo de divisiones, número mínimo de observaciones en nodo hoja y el número mínimo de ejemplos en 1 rama. Se determina generar las siguientes versiones de los árboles de decisión, con los valores recogidos en la siguiente tabla:

Tabla 13. Valores utilizados para obtener los tres arboles de decisión para el conjunto de datos Qsar: número máximo de divisiones, número mínimo de observaciones en nodo hoja y el número mínimo de ejemplos en 1 rama.

Árbol de decisión	Número máximo de divisiones	Número mínimo de observaciones en nodo hoja	Número mínimo de ejemplos en 1 rama
1	1054	1	10
2	1054	20	25
3	1054	20	40

MÉTODO DE ESTIMACIÓN DEL ERROR: VALIDACIÓN CRUZADA.

Se lleva a cabo la estimación del error del modelo mediante validación cruzada con k=5 subconjuntos disjuntos y k= 5 iteraciones, donde se destina un subconjunto como el de test para estimar el error real del modelo y los restantes para el entrenamiento/aprendizaje del modelo.

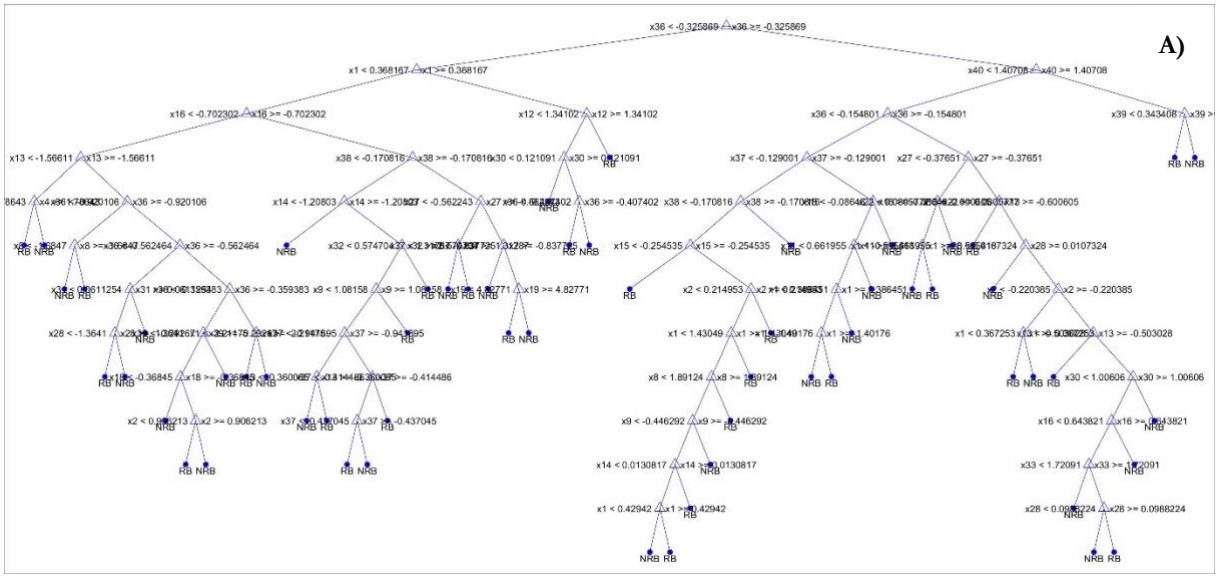
MEDIDAS DE RENDIMIENTO.

Se han calculado las mismas medidas de rendimiento que en el caso del conjunto de datos de Iris. A continuación, se muestran los resultados correspondientes, que nos servirán, junto con el árbol obtenido, a decidir la mejor versión con la que llevar a cabo la posterior comparación con los modelos de discriminante lineal y cuadrático:

Tabla 14. Medidas de rendimiento para entrenamiento y test de cada árbol obtenido para el conjunto de datos Qsar.

	Entrenamiento			Test		
	1	2	3	1	2	3
Sensibilidad (Recall)	0,9485	0,8461	0,8345	0,7896	0,7873	0,7894
Especificidad	0,9485	0,8461	0,8345	0,7896	0,7873	0,7894
Precisión	0,9529	0,8528	0,8453	0,7951	0,7934	0,7927
Valor predictivo negativo	0,9529	0,8528	0,8453	0,7951	0,7934	0,7927
Exactitud (accuracy)	0,9559	0,8661	0,8581	0,8152	0,8142	0,8142
F1-Score	0,9505	0,8489	0,8388	0,7914	0,7896	0,7908

Lo resultados obtenidos de las medidas de rendimiento (Tabla 14) muestran que no existe demasiada diferencia entre el segundo y tercer árbol. Sin embargo, el árbol obtenido por defecto demuestra que no es demasiado bueno para realizar la clasificación de nuevos datos, dadas las diferencias entre las medidas de entrenamiento y test. Por tanto, para determinar cuál es el más adecuado, hay que valorar los resultados gráficos alcanzados para cada versión.



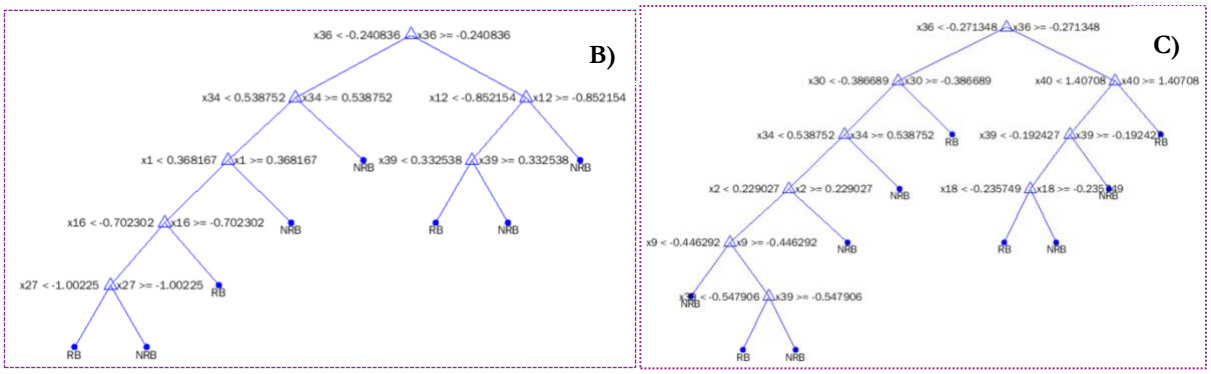


Figura 6. Representación de los tres árboles de decisión obtenidos para el conjunto de datos Qsar. **A)** Árbol con una única observación en el nodo hoja y 10 en ramas (valores por defecto). **B)** Árbol con 20 observaciones en nodo hoja y 25 en ramas. **C)** Árbol con 20 observaciones en nodo hoja y 40 en ramas.

La Figura 6.A muestra el resultado del árbol generado con los parámetros por defecto. Se puede apreciar claramente un considerable número de reglas y nodos que implican mayor complejidad y redundancia. Por el contrario, con las versiones de 20 observaciones mínimas en nodo hoja y 25 en ramas (Figura 6.B), así como con la de 20 observaciones mínimas en nodo hoja y 40 en ramas (Figura 6.C), se simplifican notablemente los árboles generados.

Por tanto, en este caso se opta por la versión 2 de árbol para llevar a cabo la comparación múltiple posteriormente ya que presenta unas medidas de rendimiento similares a la versión 3 pero un número menor de ramificaciones, es decir, es un árbol más simple.

COMPARACIÓN DE MODELOS.

SELECCIÓN DE MEDIDA DE RENDIMIENTO PARA EL TEST ESTADÍSTICO.

Se emplean las medidas del F1-Score correspondientes a test para la realización del test estadístico de los modelos.

RESULTADOS PARA LOS MODELOS IMPLICADOS.

Las medias y correspondientes desviaciones típicas para las medidas del F1-Score empleadas en el test estadístico presentan los siguientes valores:

Tabla 15. Media y desviación típica de la medida de rendimiento F1-Score para comparación de modelos lineal, cuadrático y árbol de decisión en el conjunto de datos Qsar.

Modelo	F1-Score (media ± desviación típica)
Discriminante lineal	0,8374 ± 0,0770
Discriminante cuadrático	0,7768 ± 0,0567
Árbol de decisión	0,7896 ± 0,0385

De nuevo, se mantiene un valor crítico de 0.10 para el test estadístico, con el fin de determinar si las medidas de rendimiento en los modelos resultan o no estadísticamente significativas. A continuación, se muestra la correspondiente tabla resultante del ANOVA de 1 vía:

Tabla 16. Tabla ANOVA de F1-Score del conjunto de datos Qsar.

Source	SS	df	MS	F	Prob>F
Columns	0.00834	2	0.00417	5.54	0.0198
Error	0.00903	12	0.00075		
Total	0.01736	14			

El p-valor resultante es inferior al valor crítico de 0.10 establecido, por lo que se rechaza la hipótesis nula de ausencia de diferencias estadísticamente significativas entre los modelos de discriminante lineal y cuadrático.

DIAGRAMA DE CAJAS Y BIGOTES.

La existencia de diferencias estadísticamente significativas se verifica posteriormente con la representación del diagrama de cajas y bigotes para el F1-Score sobre el conjunto de test como medida para la comparación de modelos:

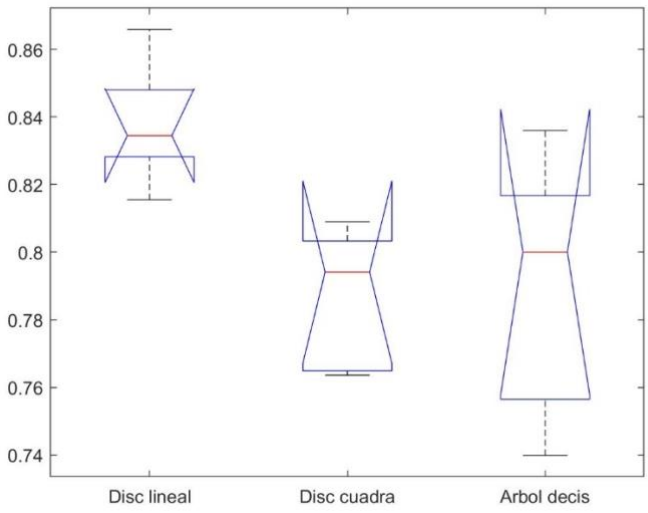


Figura 7. Diagrama de cajas y bigotes para F1-Score del conjunto de datos Qsar.

Del diagrama, se distingue que tanto el modelo de discriminante lineal como el cuadrático presentan un menor rango intercuartílico en comparación con el árbol de decisión.

TEST DE COMPARACIÓN MÚLTIPLE.

La existencia de diferencias estadísticamente significativas entre los diferentes modelos nos conduce a la realización de un test de comparación múltiple, como una representación complementaria para detectar entre los modelos que se producen las diferencias significativas:

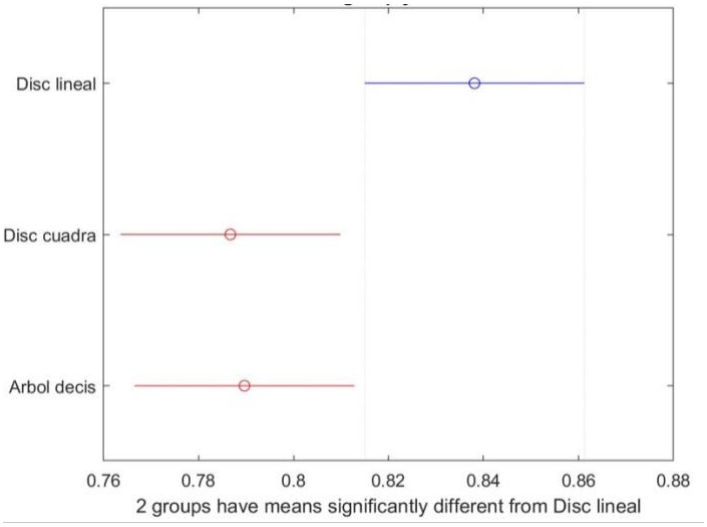


Figura 8. Representación del test de comparación múltiple para el conjunto de datos Qsar.

Este nos muestra que las diferencias estadísticamente significativas se producen entre el modelo de discriminante lineal con el cuadrático y el árbol de decisión elegido. Por el contrario, no se producen diferencias al comparar el modelo de discriminante cuadrático y el árbol de decisión.

DISCUSIÓN Y CONCLUSIONES.

Para el conjunto de QSAR del UCI Machine Learning Repository se han planteado diferentes versiones de árboles de decisión. Al igual que en el caso de Iris, en base a las medidas de rendimiento para entrenamiento y test, complementando con el gráfico resultante, se ha determinado la versión con 20 observaciones mínimas en nodo hoja y 25 en ramas. La comparación de los modelos empleando las medidas de F1-Score para test pone de manifiesto la existencia de diferencias estadísticamente significativas entre el modelo de discriminante lineal con el cuadrático y el árbol de decisión elegido.

Por tanto, debido a que las medidas de rendimiento para el árbol elegido se reducen notablemente en comparación con los otros, se valoraría el uso del discriminante lineal para el proceso de clasificación de los compuestos químicos en el conjunto. Sin embargo, podríamos considerar el árbol como un modelo complementario generando un mayor número de versiones y reducción de dimensionalidad, pues nos permitiría lograr un resultado comprensible y simplificado para clasificación.