Universidad de Valladolid

Detección de defectos en tiempo real en una línea de fabricación de tableros mediante métodos multivariantes de clasificación

Trabajo de Fin de Grado

Grado de Estadística

Universidad de Valladolid

Autor: Alejandro Rodríguez Collado

Tutor: Miguel Alejandro Fernández Temprano

Resumen

Las técnicas de análisis de datos han ido introduciéndose en el mundo de la factorías dando lugar a la Industria 4.0. En el caso del grupo Sonae Arauco diversos dispositivos les han permitido ser galardonados con premios por innovación tecnológica. Uno de sus sistemas es el "Smart Eyes", un sistema de detección de imperfecciones en los tableros producidos con filtros de imágenes.

Este proyecto tiene como objetivo crear un sistema de detección de estas imperfecciones empleando técnicas de análisis de datos y aprendizaje automático. Este proyecto es un Trabajo de Fin de Grado del Programa de Estudio Conjuntos INDat - Ingeniería Informática + Estadística. Por lo tanto constará de dos memorias, una por grado. Esta es la memoria de Estadística, y consta de las siguientes partes.

Una primera parte describirá el problema en su conjunto: datos obtenido de la empresa, características de estos y posibles variables a extraer para el posterior análisis de datos.

La segunda parte consistirá en el empleo de distintos tipos de clasificadores, principalmente distintos tipos de discriminante, árboles de decisión y extensiones de estos. Esta parte aparecerá desarrollada de forma exclusiva en la memoria del grado en Estadística y tendrá su contra-parte en la memoria del TFG de Ingeniería Informática sobre el uso de clasificadores relacionados con redes neuronales y Support Vector Machines.

La cuarta parte analizará los resultados obtenidos con cada uno de los clasificadores desarrollados en esta memoria. De esta forma, podremos conocer qué metodología ha dado lugar a resultados mejores. El rendimiento de cada clasificador aparecerá reflejado en la memoria en la cual se haya explicado previamente el funcionamiento del clasificador.

Un epílogo servirá para exponer las conclusiones sacadas relacionadas con la consecución de los objetivos marcados así como para desglosar el posible trabajo futuro a realizar.

Abstract

The data analysis techniques have been gradually introduced in the world of factories, giving birth to the 4.0 Industry. Sonae Arauco has developed various devices that have allowed them to be multiple times awarded thanks to his technical innovation. They developed "Smart Eyes", an imperfection-detecting system to be used in their board production line that works using filtering.

The goal of this project is to create an imperfection-detecting system that uses data analysis techniques and Machine Learning. This End-of-Degree Project is of a double degree (INDat – double degree in Computer Engineering and Statistics), so it consists of two memories, one per degree. This is the memory corresponding to Statistics degree and it consists of several parts.

The first part will describe the general state of the problem: data obtained from the company, its characteristics and possible useful features to extract for the posterior data analysis.

The second part will describe different classifiers as well as its results. The classifiers appearing in this part will be related to discriminant analysis, decision trees and its extensions. In the Computer Engineering End-of-Degree Project memory, you can find classifiers related to neural networks and support vector machines.

The third part will analize the obtained results of each classifier. This will allow us to know what has been the method that has obtained the best results. The performance of each classifier will be written on the memory that has previously explained the theory behind the classifier.

The epilogue will serve to explain the observed conclusions related to the achivement of proposed goals as well as break down the possible future work.

Agradecimientos

Todo el trabajo y dedicación que han hecho falta para el desarrollo de este proyecto son fruto del apoyo continuo de personas que han permanecido a mi lado desde el comienzo del mismo.

En primer lugar, querría agradecer la ayuda continua, atenta, sensata y paciente de mis tutores Arancha, Carlos y Miguel. No hubiera sido posible finalizar este trabajo sin vuestra guía y apoyo que, por otro lado, me ha brindado la oportunidad de aprender enormemente de vuestro profundo conocimiento en las materias que hemos tratado.

Muchas gracias a mis padres, Belén y José, por vuestro cariño y ayuda en el recorrido que ha sido el desarrollo del proyecto. Porque sé que, aún no entendiendo en profundidad los detalles de mi TFG, os lo habéis leído y ojeado en más de una ocasión.

Querría agradecer también el apoyo que me ha dado Alfon. Buena compañía ha sido el mejor remedio a tantos momentos en los que, al dejar de trabajar en el proyecto, me desazonaba al pensar en la larga lista de quehaceres.

Por último y no por ello menos importante me gustaría dar las gracias a mi hermana Lucía, abuelo Jesús y abuela Julia y a mis amigos, especialmente a los miembros de Teatro Horizon. Todos me habéis ayudado de alguna forma en estos meses de trabajo.

Índice general

Capítulo 1

Introducción

Este proyecto surge como colaboración con Sonae Arauco, empresa internacional de la comercialización de tablero de fibra de densidad media. El desarrollo de mis prácticas curriculares en la empresa en el primer semestre de 2017 me permitió familiarizarme con su modo de trabajo y su tecnología. Para más información respecto a los antecedentes del proyecto, véase la memoria de ingeniería informática [1].

1.1. Descripción de los datos facilitados por la empresa

1.1.1. Glosario de defectos

En esta sección describiré las características de los datos recibidos por parte de la empresa. Para más detalles, consulté la memoria del Trabajo de Fin de Grado de Ingeniería Informática [1]. Los datos recibidos por parte de la empresa se componen principalmente de dos conjuntos de datos: el Corpus I y el Corpus II.

Los defectos observados en el conjunto de datos son de tres tipos: negros, blancos y topográficos.

- **Defectos negros**: son cerca del 85 % de los defectos que se dan en la superficie de los tableros.

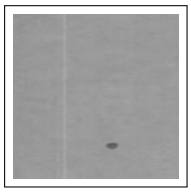

Figura 1.1: Defecto negro de tamaño pequeño.

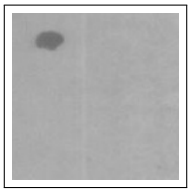

Figura 1.2: Defecto negro grande.

- **Defectos blancos**: aparecen debido a la aparición de restos de fibras en la superficie del tablero. Suponen cerca 6 % de los tableros que aparecen en las muestras entregadas. Necesitan de una iluminación homogénea y adecuada para que se detecten adecuadamente. Las Figuras 1.3 y 1.4 muestran ejemplos de este tipo de defecto.

Figura 1.3: Fibra de tamaño pequeño.

Figura 1.4: Fibra de tamaño grande.

- **Defectos Topográficos**: denominados de esta forma debido a que están relacionados con contornos o relieves. Se les puede considerar defectos negros y blancos a la vez debido a las zonas más claras y más óscuras que aparecen. Son muy diversos, y aparecen en el 9 % de las taras halladas en las imágenes de las muestras. Algunos ejemplos de defectos topográficos son:

- Hendiduras.

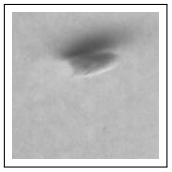

Figura 1.5: Defecto cuyo contorno es más oscuro.

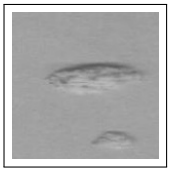

Figura 1.6: Defecto de contorno más claro.

- Grietas de separación y de unión.

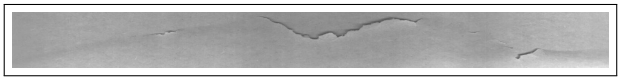

Figura 1.7: Imagen de una grieta de separación: resquebrajamiento simple.

Figura 1.8: Imagen de una grieta de unión.

- Roturas.

Figura 1.9: Imagen de una rotura en el lateral derecho del tablero.

Figura 1.10: Imagen de una rotura en el lateral izquierdo del tablero.

- Humedades.

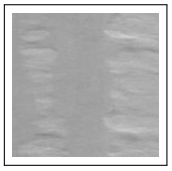

Figura 1.11: Rastro de agua en la superficie del tablero.

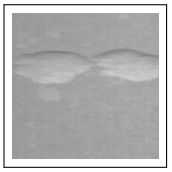

Figura 1.12: Gotas de agua en el tablero.

- Bolsas de aire.

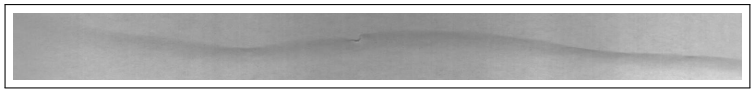

Figura 1.13: Bolsa de aire acompañada de una pequeña grieta.

Figura 1.14: Bolsa de aire cuya detección es más compleja.

Para obtener más información acerca de los diferentes tipos de defectos, véase el glosario de la memoria de ingeniería informática [1].

1.1.2. Corpus de imágenes

La empresa nos ha facilitado dos conjuntos de imágenes para la realización de la investigación científica. Cada uno de ellos tiene diferentes características y naturaleza. La distribución de clases de estos Corpus aparece en la Tabla 1.1:

Conjunto de imágenes empleado			
	Corpus I	Corpus II	Total
Sin defectos	709	3893	4602
Con defectos	6458	107	6565
Total	7167	4000	11167

Tabla 1.1: Distribución de observaciones en el total de imágenes empleadas.

Empleando la clasificación de defectos desarrollada en el glosario, los defectos les podemos deglosar en (Tabla 1.2):

Conjunto de imágenes empleado			
Defecto	Corpus I	Corpus II	Total
Negro	5497	91	5588
Blanco	546	15	561
Topográfico: Bolsas de aire	41	0	41
Topográfico: Gotas de agua	17	0	17
Topográfico: Grietas de separación	64	0	64
Topográfico: Grietas de unión	46	0	46
Topográfico: Hendiduras	94	1	95
Topográfico: Rotura horizontal	66	0	66
Topográfico: Rotura lateral	71	0	71
Topográfico: Surcos de Agua	55	0	55
Total	6458	107	6565

Tabla 1.2: Distribución de los defectos en el conjunto total de imágenes.

Los totales no son la adición del total del observaciones de cada tipo de defecto, ya que algunas imágenes tienen presentes más de un tipo de defecto.

Las dos siguientes subsecciones describen las propiedades y composición de los dos Corpus de imágenes empleados.

1.1.2.1. Corpus I: Imágenes "defectuosas"

El Corpus I de imágenes se compone exclusivamente de imágenes clasificadas como defectuosas por el sistema basado en un filtrado pesado que tienen en funcionamiento en la empresa ('Smart Eyes' [1]). En la toma de las imágenes se han mantenido fijos los parámetros relativos a la misma, auqnue los datos recogidos no corresponden como tal a ningún marco de muestreo establecido de antemano.

Figura 1.15: Ejemplo de imagen 'original' del Corpus I.

La distribución de clases del Corpus I aparece en la Tabla 1.3 y la clasificación en la tipología de defectos desarrollada en el glosario en la Tabla 1.4.

Corpus I	
	Observaciones
Sin defectos	709
Con defectos	6458
Total	7167

Tabla 1.3: Distribución de observaciones del Corpus I.

Corpus I	
Defecto	Observaciones
Negro	5497
Blanco	546
Topográfico: Bolsas de aire	41
Topográfico: Gotas de agua	17
Topográfico: Grietas de separación	64
Topográfico: Grietas de unión	46
Topográfico: Hendiduras	94
Topográfico: Rotura horizontal	66
Topográfico: Rotura lateral	71
Topográfico: Surcos de Agua	55

Tabla 1.4: Distribución de defectos observados en el Corpus I.

1.1.2.2. Corpus II: Imágenes "sin defectos"

Este Corpus se recibió en forma de cuatro vídeos, al contrario del formato de instantáneas del Corpus I. Esto da lugar a imágenes cuya resolución es mucho más pobre además de necesitar de un recorte manual y una gama de grises mucho más oscura. Por otro lado, no conocemos el comportamiento de Smart Eyes sobre el mismo ya que se ha tenido que desactivar para poder realizar la grabación.

La distribución de procedencia y clase de las imágenes que conforman este Corpus aparece en la Tabla 1.5. La distribución de etiquetas de las imágenes con defectos se encuentra en la Tabla 1.6:

Corpus II					
Cámara:	Izquierda	Centro-izquierda	Centro-derecha	Derecha	Total
Sin defectos	934	1069	841	1049	3893
Con defectos	11	9	48	39	107
Total	945	1078	889	1088	4000

Tabla 1.5: Distribución de clases observadas en el Corpus II.

Corpus II					
Defecto	Izquierda	Centro-izquierda	Centro-derecha	Derecha	Total
Negro	8	9	43	31	91
Blanco	3	0	4	8	15
Topográfico: Hendiduras	0	0	1	0	1

Tabla 1.6: Distribución de defectos observados en el Corpus II.

1.2. Características extraídas

Las variables de entrada a los clasificadores se pueden agrupar en tres grupos:

- Áreas de error relativo en negro y blanco.
- Caracterizaciones de las distribuciones de grises.
- Componentes de la transformada discreta del coseno (DCT).

1.2.1. Áreas de error relativo en negro y blanco

Variables básicas calculadas de forma sencilla sobre la salida del proceso de filtrado de imágenes especificado en la memoria de ingeniería informática [1]. Son calculadas como la adición de los píxeles marcados como potencialmente defectuosos tras el procedimiento de filtrado entre el área estimado de la imagen. Se calculan por separado para los procedimientos de filtrado de defectos negros y blancos. La Figura 1.16 refleja la salida básica del procedimiento empleado para el cálculo de las áreas de error relativas a partir de las imágenes filtradas y binarizadas.

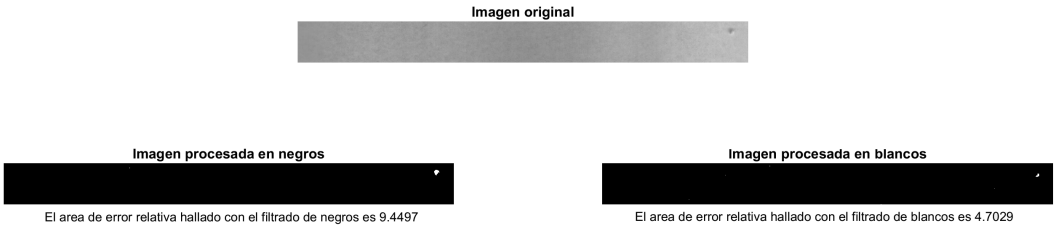

Figura 1.16: Imagen acompañada de las resultantes del filtrado de negros y blancos, así como el área de error relativa asociada estimada. El valor ha sido multiplicado por un factor de escala.

1.2.2. Caracterizaciones de las distribuciones de grises

Esta colección de variables busca resumir la distribución de grises de las imágenes con las que se ha trabajado. Esta es la frecuencia observada relativa de cada uno de los tonos de gris en la imagen. Hay 256 tonos de gris codificados entre el 0 y el 255, de tal forma que el 0 corresponde al color negro y el 255 al negro. Una representación habitual de esta distribución es el histograma de grises (Figura 1.17).

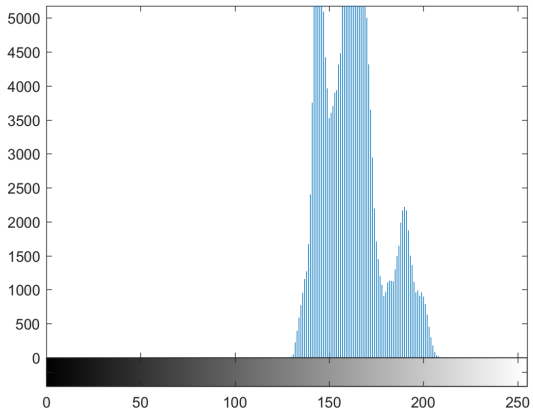

Figura 1.17: Histograma de grises asociado a la distribución de grises de una imagen en bruto.

La información subyacente a esta distribución que genera 256 variables se ha intentado sintetizar con dos caracterizaciones diferentes. Estas conformarán los dos conjuntos de atributos con los que se trabajara simultáneamente. Las caracterizaciones empleadas han sido:

- Conjunto A - PCA: uso de componentes principales [2], de tal forma que la reducción de dimensionalidad es muy notoria.

- Conjunto B - Estadísticos: uso de estadísticos característicos (media, varianza, asimetría, kurtosis, mediana, cuartil-10 % y cuartil-90 %) para una reducción menor de la dimensionalidad en la que podría existir más información útil para discriminar.

Se han extraído estas variables de la imagen original y de las filtradas sin que se hayan binarizado.

1.2.3. Transformada Discreta del Coseno (DCT)

El último subconjunto de variables estudiado son las componentes de la transformada discreta del coseno (DCT [3]). Es una técnica según la cual se expresa información como adición de funciones cosenoidales oscilando a diferentes frecuencias. Es similar a la transformada de Fourier (DFT), solo que esta emplea también funciones senoidales.

$$x[n] = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} w[k] C_x[k] \cos\left(\frac{\pi}{2N} k(2n+1)\right) & \text{si } 0 \leq n < N \\ 0 & \text{en otro caso} \end{cases}$$

$$w[k] = \begin{cases} \frac{1}{2} & \text{si } k = 0 \\ 1 & \text{si } 0 \leq k < N \end{cases}$$

Los coeficientes de estas funciones coseno se calculan de la siguiente forma a partir de los valores de $x[n]$:

$$C_x[k] = \begin{cases} \sum_{n=0}^{N-1} 2x[n] \cos\left(\frac{\pi}{2N} k(2n+1)\right) & \text{si } 0 \leq k < N \\ 0 & \text{en otro caso} \end{cases}$$

Se han aplicado dos transformaciones sucesivas a las componentes extraídas de la DCT: se han elevado al cuadrado (haciendo positivos los valores en su totalidad) y se ha aplicado después el logaritmo (así la distribución subyacente es más suave y cumple de forma más fuerte las hipótesis de normalidad). En la imagen filtrada binarizada y sin binarizar, se han extraído los 30 primeros términos de la DCT, incrementando el número de atributos de los conjuntos de variables en 60.

La Figura 1.18 muestra un esquema de las imágenes con las que se ha trabajado así como las características extraídas de cada una de ellas.

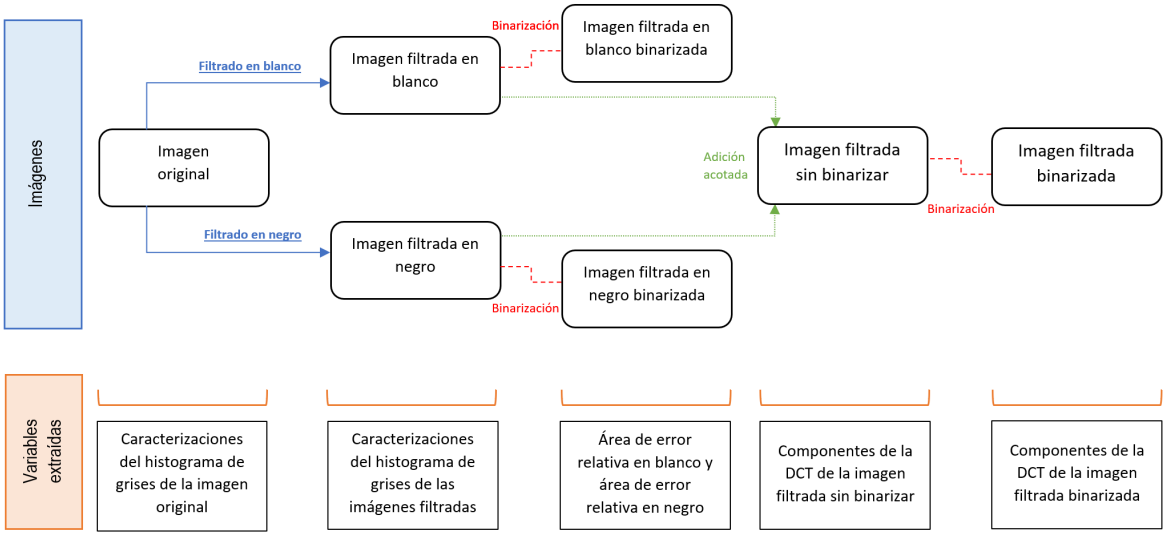

Figura 1.18: Representación esquemática de las imágenes con las que se ha trabajado y variables que se extraen en cada fase.

1.3. Conjuntos de variables predictoras consideradas

Esta subsección detalla las variables que componen los conjuntos de atributos que servirán de entrada a los sistemas de análisis supervisado considerados. Su principal diferencia radicará en el uso de las componentes principales (Conjunto A) o los estadísticos (Conjunto B) como caracterización de la distribución de grises, mientras que las áreas de error relativas y las componentes DCT extraídas serán variables comunes.

Se estudiaron otras alternativas como un conjunto de variables C compuesto unicamente por las áreas de error relativa y las componentes de la DCT o alternativas a los conjuntos A y B que no empleasen la información de la DCT. Sin embargo, los resultados siempre resultaron ser más pobres lo que hizo que se descartase trabajar con ellos en las primeras fases del proyecto.

Conjunto A: uso de PCA
(2 + 8 + 60 variables)

Las variables que componen este conjunto de datos son:

- Área de error relativa en la imagen procesada en negros y en la imagen procesada en blancos. Tomarán valores naturalmente bajos entre 0 y 1 (2 variables).
- A través del análisis de componentes principales extraemos:
  - Componentes 1-4 de la distribución de grises de la imagen original. En las imágenes del Corpus I, cuatro componentes recogen una inercia del 80.20 %. El resto de componentes son residuales. En el Corpus II la inercia alcanza el valor de 88.53 % con 4 componentes. (4 variables).
  - Componentes 1-2 de la distribución de grises de la imagen tratada en blancos. La varianza explicada con 2 componentes en el Corpus I es del 96.91 % mientras que en el Corpus II es del 97.46 %. (2 variables).
  - Componentes 1-2 de la distribución de grises de la imagen tratada en negros. La inercia recogida por dos componentes en el caso del Corpus I ha sido del 96.30 %. En el Corpus II este valor es del 97.13 % (2 variables).
- DCT de la imagen tratada sin binarizar y binarizada (30 × 2 variables)

La Tabla 1.7 refleja aquellas variables que forman parte de este conjunto de datos.

Variables del conjunto de variables A (70)					
	Area de Error (2)	AreaErrorB	AreaErrorN		
PCA	Img. original (4)	CompPrinOr1	CompPrinOr2	CompPrinOr3	CompPrinOr4
	Img. filtrada en blanco (2)	CompPrinTratB1	CompPrinTratB2		
	Img. filtrada en negro (2)	CompPrinTratN1	CompPrinTratN2		
DCT	Img. filtrada sin binarizar (30)	Componentes de la 1 a la 30 extraídas.			
	Img. filtrada binarizada (30)	Componentes de la 1 a la 30 extraídas.			

Tabla 1.7: Variables del conjunto de datos A.

Conjunto B: uso de estadísticos
(2 + 21 + 60 variables)

Las siguientes variables componen este conjunto de variables:

- Área de error relativa en la imagen procesada en negros y en la imagen procesada en blancos. Tomarán valores naturalmente bajos entre 0 y 1 (2 variables).
- Media, varianza, asimetría (skewness), kurtosis, mediana, cuartil-10 % y cuartil-90 % de las distribuciones de grises de la imagen original, tratada en blancos y tratada en negros. Serán valores entre 0 y 255 (7 × 3 variables).

- DCT de la imagen tratada sin binarizar y binarizada (30 × 2 variables)

La Tabla 1.8 resume aquellos atributos que forman parte de este conjunto de variables.

Variables del conjunto de variables B (83)					
	Area de Error (2)	AreaErrorB	AreaErrorN		
Estadísticos	Img. original (7)	MediaOr	VarianzaOr	SkewnessOr	KurtosisOr
		MedianaOr	P10Or	P90Or	
	Img. filtrada en blanco (7)	MediaTratadaB	VarianzaTratadaB	SkewnessTratadaB	KurtosisTratadaB
		MedianaTratadaB	P10TratadaB	P90TratadaB	
	Img. filtrada en negro (7)	MediaTratadaN	VarianzaTratadaN	SkewnessTratadaN	KurtosisTratadaN
		MedianaTratadaN	P10TratadaN	P90TratadaN	
DCT	Img. filtrada sin binarizar (30)	Componentes de la 1 a la 30 extraídas.			
	Img. filtrada binarizada (30)	Componentes de la 1 a la 30 extraídas.			

Tabla 1.8: Variables del conjunto de datos B.

Capítulo 2

Metodología

En este capítulo se expondrán cada uno de los modelos desarrollados en el proyecto y que posteriormente serán evaluados en el Capítulo de los resultados (Capítulo 3).

La línea de trabajo seguida es la siguiente: se comenzará con un sencillo sistema de referencia que dará paso a clasificadores básicos como el discriminante lineal. Posteriormente se desarrollarán diferentes tipos de árboles de decisión para dar paso finalmente a la agregación de estos en forma del bagging y boosting. Por lo tanto, la complejidad de los clasificadores aumentará gradualmente.

De cada uno de los clasificadores se hará una breve contextualización de su origen. Posteriormente se desarrollará el fundamento formal de la técnica así como posibles aplicaciones prácticas ya implementadas en las que el procedimiento en cuestión ha logrado resultados altamente satisfactorios. A mayores, se desarrolla un apartado sobre la implementación utilizada con cada uno de los procedimientos.

Sobre cada uno de los distintos tipos de clasificadores se estudiará el área bajo la curva ROC asociada, la tasa de equierror, la sensibilidad, la especificidad y precisión conseguidas con los siguientes modelos:

- **Modelo A**: desarrollo de un único clasificador (imagen con defecto-sin defecto) con el conjunto de variables A (PCA). Se trata de un conjunto de variables más reducido en dimensionalidad que el B.
- **Modelo B**: desarrollo de un único clasificador (imagen con defecto-sin defecto) con el conjunto de variables B (Estadísticos). Se trata de un conjunto con más variables que el A que podrían aportar una mayor varianza explicada.

En cada uno de los clasificadores se hará una selección de características o un estudio de la relevancia de las variables empleadas. La selección de modelos óptimos se realizará utilizando como criterio la tasa de acierto (2.0.1.1), a excepción de que el modelo no sea balanceado en cuanto a distribución de clases y tenga una fuerte preferencia por una de ellas. En tal caso la medida de referencia será el F-Score (2.0.1.2), media armónica de la sensibilidad y la especificidad, que permitirá la elección teniendo en cuenta tanto la clasificación de imágenes con defectos como las que carecen de ellos.

En los casos de clasificadores cuya salida sea probabilística o acotada se expondrán las curvas ROC (2.0.1.4), tasa de equierror (2.0.1.3), tasa de acierto y la matriz de confusión (2.0.1) de frecuencias relativas de los modelos estimadas mediante validación cruzada de 10 particiones. En caso contrario, se expondrán tan solo la tasa de acierto estimada y la matriz de confusión de frecuencias relativas aproximadas mediante validación cruzada (10 particiones).

2.0.1. Definiciones previas

Para el estudio de resultados así como en la selección de modelos óptimos se han utilizado una serie de conceptos definidos a partir de la matriz dos por dos básica que, en nuestro caso, corresponde a una matriz de confusión de clasificación binaria.

En la Tabla 2.1 vemos un ejemplo de matriz de confusión dos por dos [4]. En horizontal se define la clase verdadera de la observación, mientras que en vertical aparece la clase predicha o etiqueta.

		Clase	
		Sin defecto	Con defecto
Etiq.	Sin defecto	TN	FN
	Con defecto	FP	TP

Tabla 2.1: Matriz de confusión dos por dos.

En la diagonal de esta matriz aparecen las observaciones bien clasificadas: los verdaderos positivos (True Positive - TP), en nuestro caso las imágenes con defectos clasificadas como con defectos, y los verdaderos negativos (True Negative - TN), imágenes sin defectos clasificadas como tal. En la posición inferior izquierda están los tableros sin defectos clasificados como si los tuviese (falsos positivos, False Positive - FP) y en la posición derecha superior los tableros con defectos clasificados como sino no los tuviese (falsos negativos, False Negative - FN).

Llamamos tasa de verdaderos positivos o sensibilidad (True Positive Rate - TPR) a la probabilidad existente de que una observación, condicionada a que tenga defectos, sea clasificada como tal.

$$\text{Sensibilidad} = \text{TPR} = P(\text{Etiqueta=Defectuoso}|\text{Clase=Defectuoso}) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

De forma análoga definimos la tasa de verdaderos negativos o especificidad (True Negative Rate - TNR) como la probabilidad existente de que una observación, condicionada a que no tenga defectos, sea clasificada correctamente como que no los tuviese.

$$\text{Especificidad} = \text{TNR} = P(\text{Etiqueta=Sin defectos}|\text{Clase=Sin defectos}) = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

A mayores podemos definir otras dos tasas relacionadas con los errores a la hora de clasificar del sistema. La tasa de falsos negativos (False Negative Rate - FNR) es la probabilidad complementaria a la sensibilidad. Se puede definir como la probabilidad existente de que una observación, condicionada a que tenga defectos, sea clasificada como si no los tuviese.

$$\text{FNR} = P(\text{Etiqueta=Sin defectos}|\text{Clase=Defectuoso}) = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

La tasa de falsos positivos (False Positive Rate - FPR) es la probabilidad existente de que una observación, condicionada a que no tenga defectos, sea clasificada como si los tuviese. Se trata de la probabilidad complementaria a la especificidad.

$$\text{FPR} = P(\text{Etiqueta=Defectuoso}|\text{Clase=Sin defectos}) = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

2.0.1.1. Tasa de acierto

La tasa de acierto o precisión es una medida de evaluación de sistemas de referencia [5]. Se define como la proporción de observaciones bien clasificadas entre el total de las mismas. En base a las posiciones definidas en la matriz dos por dos básica, podemos definir la tasa de acierto:

$$\text{Tasa de acierto} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

La tasa de error es la probabilidad complementaria a la tasa de acierto y se calcula como la proporción de observaciones mal clasificadas entre el total de las mismas. Se emplean estimaciones de ésta para seleccionar modelos ante un conjunto de los mismos, siempre y cuando los modelos obtenidos sean balanceados clasificando en las dos clases. De no darse esto, se pueden emplear medidas relacionadas con la especificidad y sensibilidad como es el F-Score, definido más adelante en la Subsección 2.0.1.2.

Para la estimación de la verdadera tasa de acierto de un modelo se suele emplear un conjunto de datos test sobre el que se evalúa la misma. Ante la falta de datos suficientes para hacer esto, se emplean diversas técnicas basadas en repeticiones de muestreos aleatorios como es el Leave-One-Out (LOO) o la validación cruzada de k particiones [6].

En los resultados relativos a la tasa de acierto se ha aportado a mayores un intervalo de confianza. Ante la falta de independencia de las particiones de la validación cruzada, el intervalo se basa en una distribución t de Student con 9 (k-1) grados de libertad. Si llamamos $\bar{e}$ a la tasa de acierto de acierto media de las particiones, S_e^2 al estimador de la varianza muestral y $t_{\alpha/2,9}$ al valor de una distribución t de Student que deja a su derecha una probabilidad $^{\alpha}/_{2}$, la expresión del intervalo de confianza $1-\alpha$ será:

$$\bar{e} \mp t_{\alpha/2,9} \frac{S_e}{\sqrt{10}}$$

2.0.1.2. F-Score

El F-Score es una medida de precisión para tests y clasificadores empleada especialmente en casos donde la tasa de acierto no aporta información precisa. Esto se deberá al desbalance entre clases en el conjunto de entrenamiento: clasificar solo a la clase mayoritaria dará lugar a una tasa de acierto elevada y sin embargo la capacidad clasificadora del clasificador será nula. Esto es sinónimo de que o la sensibilidad o la especificidad alcanzan un valor muy bajo.

La definición empleada del F-Score es la media armónica de la sensibilidad y especificidad. La expresión analítica es la siguiente:

$$F_1 = \frac{2}{\frac{1}{\text{Sensibilidad}} + \frac{1}{\text{Especificidad}}} = 2 \cdot \frac{\text{Sensibilidad} \cdot \text{Especificidad}}{\text{Sensibilidad} + \text{Especificidad}}$$

2.0.1.3. Tasa de Equierror

La tasa de equierror (EER - Equal Error Rate) es una medida de evaluación de sistemas de clasificación empleado comúnmente en dispositivos de biometría. Se define como el punto en el que se igualan la tasa de falsos positivos y la tasa de falsos negativos. Estas dos tasas son probabilidades condicionadas que definimos previamente a partir de la matriz de confusión dos por dos.

El umbral que seleccionaremos será aquel que permita que se igualen la tasa de falsos positivos y falsos negativos tal y como podemos ver en la Figura 2.1.

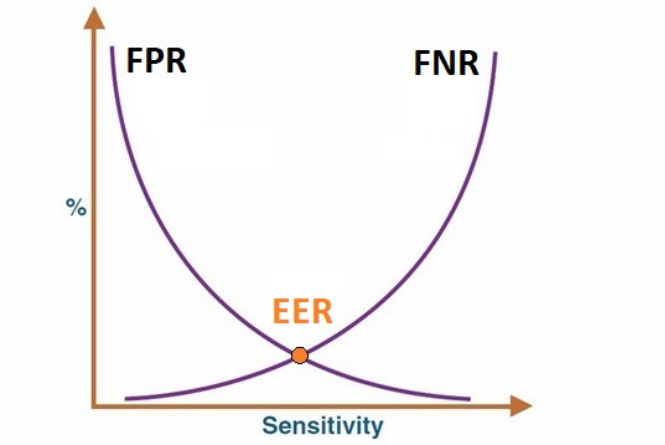

Figura 2.1: Representación gráfica de la definición de tasa de equierror (EER - Equal Error Rate). En la figura se representan la tasa de falsos positivos (FPR) y la tasa de falsos negativos (FNR) en función de la sensibilidad.

2.0.1.4. Curvas ROC (Receiver Operating Characteristic)

Las curvas ROC son representaciones gráficas de la tasa de verdaderos positivos (sensibilidad) frente a la tasa de falsos positivos (uno menos la especificidad) en un sistema de clasificación binaria según se varía el umbral que

separa entre clases [5]. Se emplea mucho para valorar la calidad de tests de diagnosis de enfermedades así como para evaluar comparativamente clasificadores [7].

La Figura 2.2 muestra lo que sería el espacio sobre el que se definen las curvas ROC.

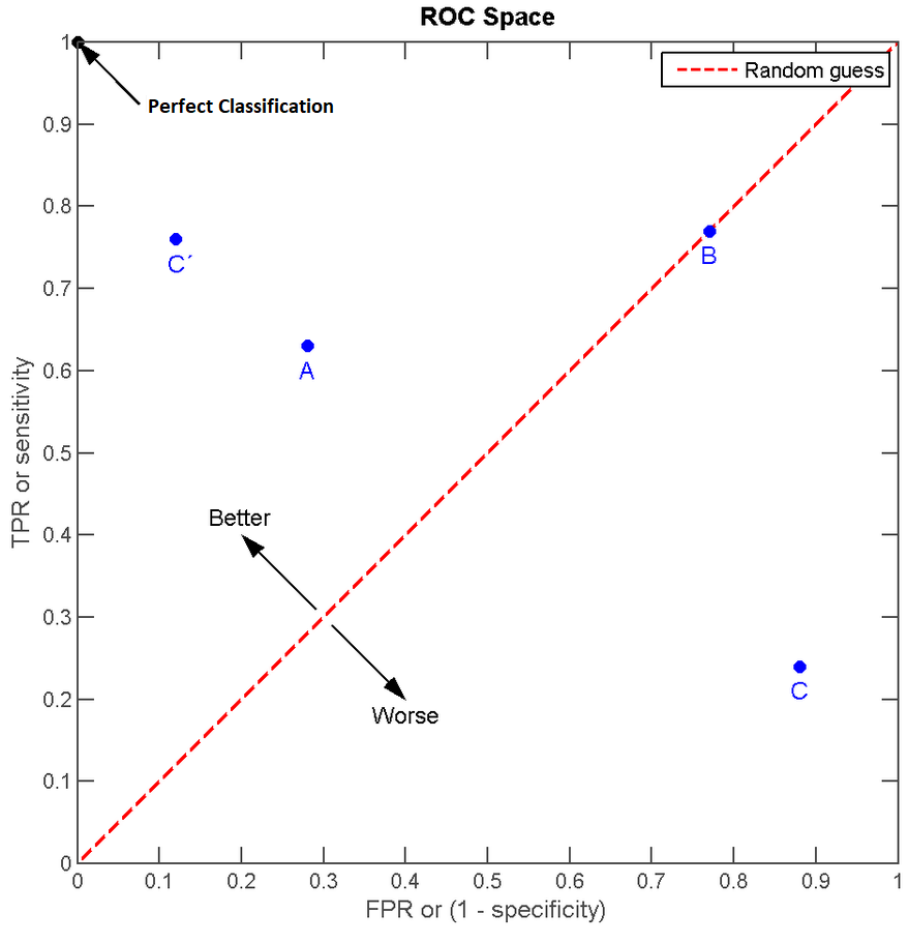

Figura 2.2: Representación gráfica del espacio de una curva ROC así como diversos puntos relevantes [8].

Se trata de una representación que cuenta con dos partes claramente diferenciadas. Esta división se da por la línea discontinua de color rojo en la Figura. Esta frontera supone la asignación al azar en un sistema de clasificación con dos clases. Los puntos por encima suponen que el clasificador es capaz de discriminar de cierta forma entre las clases mientras que si queda por debajo la asignación al azar es más efectiva. Si se alcanzase el punto (0,1), correspondiente a que tanto la sensibilidad como la especificidad sean uno, la clasificación perfecta sería posible.

Para medir la calidad de un clasificador en función de su curva ROC se emplea normalmente la AUC (Area Under Curve), que se define como el área recogido bajo la curva ROC. Su valor máximo es uno y cuanto mayor sea, mejor será la precisión del clasificador en cuestión.

Una clasificación simple para calificar el funcionamiento de un test o clasificador puede ser la que muestra la Tabla 2.2.

Cada uno de los puntos de la curva ROC representados así como el valor de la AUC se han estimado por validación cruzada de diez particiones para aproximarse al verdadero valor, empleando la media (μ_{AUC}) de estos como estimación. También se ha estimado la desviación estándar del AUC (σ_{AUC}) a partir de las diez particiones.

AUC	Calidad del clasificador
0.00-0.50	La asignación al azar es más eficaz.
0.50	Equivalente a asignar al azar.
0.50-0.60	Malo.
0.60-0.70	Pobre.
0.70-0.80	Bueno.
0.80-0.90	Muy bueno.
0.90-1.00	Excelente.

Tabla 2.2: Calidad de un sistema de clasificación según el valor de la AUC definida a partir de una curva ROC [9].

Esta ha servido para crear los intervalos de confianza del 95 % ($\alpha = 0.05$) que aparecen en las tablas de resultados del documento de la siguiente forma:

$$\mu_{AUC} \mp z_{1-\frac{0.05}{2}} \sigma_{AUC} = \mu_{AUC} \mp 1.96 \sigma_{AUC}$$

La Figura 2.3 muestra el formato de curva ROC empleado. El azar aparece representado en gris y la curva como tal está en morado. A mayores, se han representado las curvas ROC medias correspondientes al defecto blanco y al defecto negro. Éstas en las Figuras aparecerán en azul y rojo respectivamente. En la leyenda aparece indicado el valor medio del AUC y su desviación estándar para las tres curvas (defecto general, defecto blanco y defecto negro).

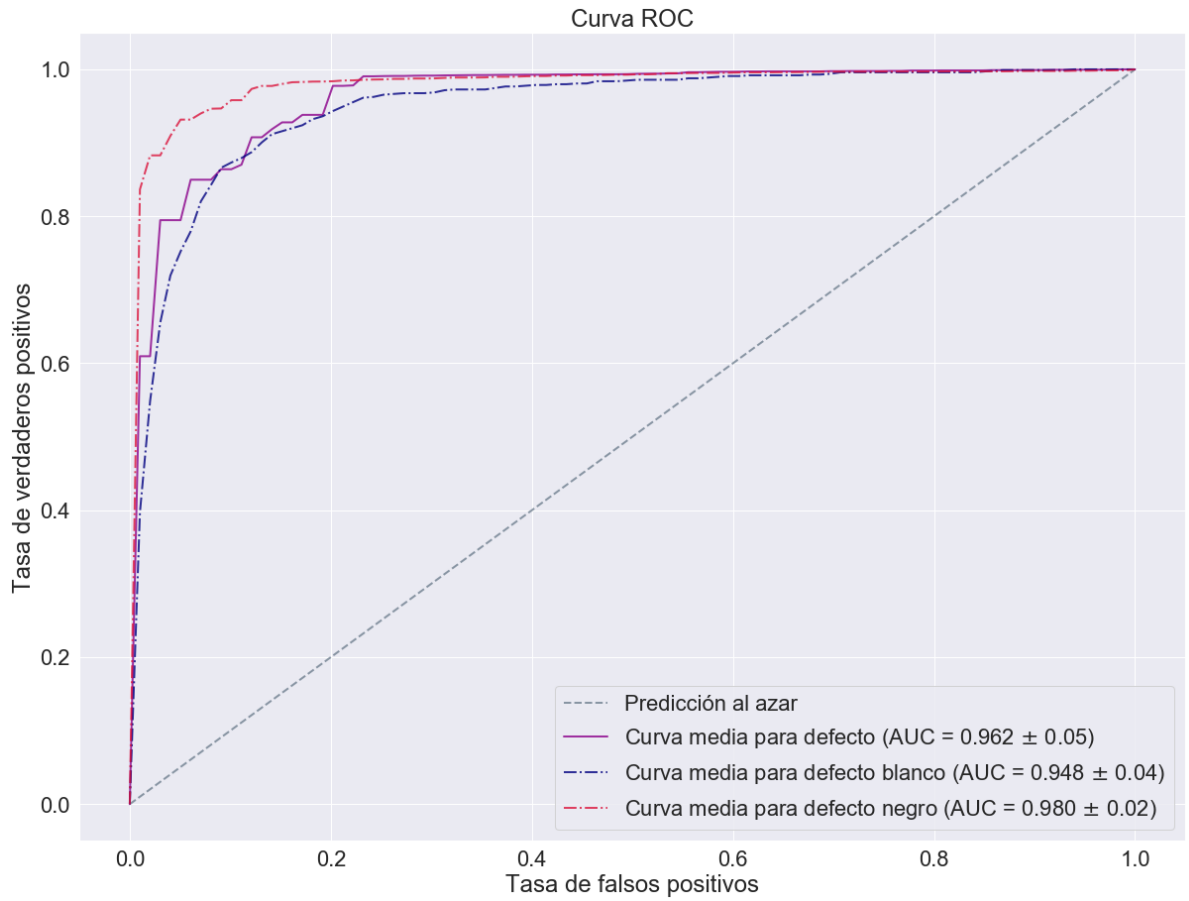

Figura 2.3: Ejemplo de curva ROC. El azar queda representado en gris y la curva como tal en morado. Las curvas ROC medias de defecto blanco y negro aparecen de color azul y rojo respectivamente.

2.1. Sistema de referencia

Con las imágenes procesadas con nuestros filtros como describe el apartado correspondiente dentro de la memoria de ingeniería informática, una primera aproximación consiste en hacer un cálculo aproximado del área de error y etiquetar como defectuosos en base a un umbral. Aquellos que superen este determinado valor umbral serán clasificados como defectuosos. Recordemos que la salida del proceso de filtrado serán dos imágenes cuyas matrices se compondrán solo de ceros y unos, correspondientes a la posible ausencia o presencia estimada de defecto respectivamente.

El área relativa se puede calcular aproximadamente de forma sencilla contando con las propiedades de la imagen (altura y anchura de la misma) así como con el total de píxeles de potenciales errores, tal y como se describió en la Sección 1.2.1.

Como estudio preliminar se desarrollarán dos posibles vías de trabajo en cuanto al sistema de referencia:

- **Dos umbrales**: se trabajará con un modelo de dos variables. Se definirán dos umbrales diferentes para clasificar en imagen con defecto negro-sin defecto negro (con el área de error relativa obtenida del procesamiento en negro) y en imagen con defecto blanco-sin defecto blanco (con el área de error relativa obtenida del procesamiento en blanco). La disyunción de ambas será la clasificación a tener en cuenta para definir la tasa de error.
- **Un umbral**: se trabajará con un modelo de una única variable. Un umbral desarrollado con la suma de áreas de error relativas halladas nos permitirá clasificar en imagen con defecto-sin defecto.

Se ha empleado para la selección del umbral la tasa de equierror (Sección 2.0.1.3). El umbral se ha seleccionado por separado en el Corpus I y en el II para seguir una línea de trabajo similar a la desarrollada con los clasificadores. Los resultados obtenidos aparecen reflejados a lo largo del correspondiente apartado en el capítulo de resultados (Sección 3.2).

2.2. Análisis discriminante

El análisis discriminante (Linear Discriminant Analysis - LDA) [10] es un método de aprendizaje supervisado en el que se busca una combinación entre variables capaz de distinguir entre dos o más categorías o tipos de individuos. Su creador fue Ronald Fisher en 1936.

Llamaremos a las categorías posibles de las observaciones y (imagen con defecto-imagen sin defecto) y la matriz de características extraídas X. La dimensión de X será de n observaciones $\times$ p variables.

$$y = \begin{cases} 0 & \text{si la imagen no tiene defectos} \\ 1 & \text{si la imagen tiene defectos} \end{cases}$$

El análisis discriminante supone cierta una hipótesis inicial: las distribuciones condicionales al grupo siguen distribuciones normales multivariantes de media μ_1 y μ_2 y matriz de covarianzas Σ_1 y Σ_2 [5].

$$X|y = 1 \sim N_p(\mu_1, \Sigma_1)$$

$$X|y = 2 \sim N_p(\mu_2, \Sigma_2)$$

Sus principales aplicaciones han estado relacionadas con el posicionamiento y el marketing de productos [11]. Ha demostrado alcanzar bajas tasas de error en la predicción de posibles insolvencias en clientes y una de sus aplicaciones más recientes es en ámbitos de reconocimiento facial. También se emplea en estudios clínicos para definir el estado médico de un paciente haciendo uso de distintas mediciones biológicas [12] o en clasificación automática de textos [13].

2.2.1. Análisis discriminante lineal (LDA)

En el caso del discriminante lineal (LDA - Linear Discriminant Analysis) se hace una suposición a mayores de homocedasticidad ($\Sigma_1 = \Sigma_2 = \Sigma$) que supone una simplificación del modelo. De esta forma la función de discriminación pasa a ser lineal. Podemos definir la función δ de la siguiente forma, siendo π_k la probabilidad a priori definida para el grupo k y f_k su función de densidad:

$$\delta_k(x) = \log(\pi_k f_k(x)) \propto -\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k) + \log(\pi_k)$$

La clasificación predicha($\hat{y}$) la podemos definir de la siguiente forma:

$$\hat{y} = \begin{cases} 0 & \text{si } \delta_2(x) - \delta_1(x) \leq 0; \text{si}(\mu_1-\mu_2)'\Sigma^{-1}(x - \frac{\mu_1+\mu_2}{2}) \leq \log(\frac{\pi_2}{\pi_1}) \\ 1 & \text{si } \delta_2(x) - \delta_1(x) > 0; \text{si}(\mu_1-\mu_2)'\Sigma^{-1}(x - \frac{\mu_1+\mu_2}{2}) > \log(\frac{\pi_2}{\pi_1}) \end{cases}$$

Este procedimiento da lugar a una combinación lineal de pesos ($w \in \Re^p$) acompañados de un peso adicional o término independiente (w_0).

$$\hat{y} = \begin{cases} 0 & \text{si } w'x + w_0 \leq 0 \\ 1 & \text{si } w'x + w_0 > 0 \end{cases}$$

La parte izquierda de la Figura 2.4 muestra claramente la linealidad resultado de la discriminación lineal en el ejemplo clásico del conjunto de datos 'Iris'.

2.2.2. Análisis discriminante cuadrático (QDA)

Las hipótesis base definidas para el discriminante cuadrático son las básicas del análisis discriminante por lo que puede haber heterocedasticidad. Con ello se admiten matrices de covarianzas distintas que darán lugar a una función cuadrática. La función de la que dependerá la clasificación predicha tomará la siguiente expresión:

$$Q(x) = \log(\pi_2 f_2(x) - \pi_1 f_1(x)) = \log(\frac{\pi_2}{\pi_1}) - \frac{1}{2}\log(\frac{\Sigma_2}{\Sigma_1}) - \frac{1}{2}((x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)) = w'x + w_0 + x'\Omega x$$

Siendo Ω una matriz $p \times p$ simétrica. La clasificación predicha ($\hat{y}$) será por lo tanto:

$$\hat{y} = \begin{cases} 0 & \text{si } w'x + w_0 + x'\Omega x \leq 0 \\ 1 & \text{si } w'x + w_0 + x'\Omega x > 0 \end{cases}$$

De esta forma se obtienen fronteras definidas por funciones cuadráticas tal y como podemos observar en la Figura 2.4. A la derecha se encuentra la frontera definida por un procedimiento de discriminante lineal en el problema clásico Iris, mientras que a la derecha queda la frontera cuadrática que define el discriminante cuadrático.

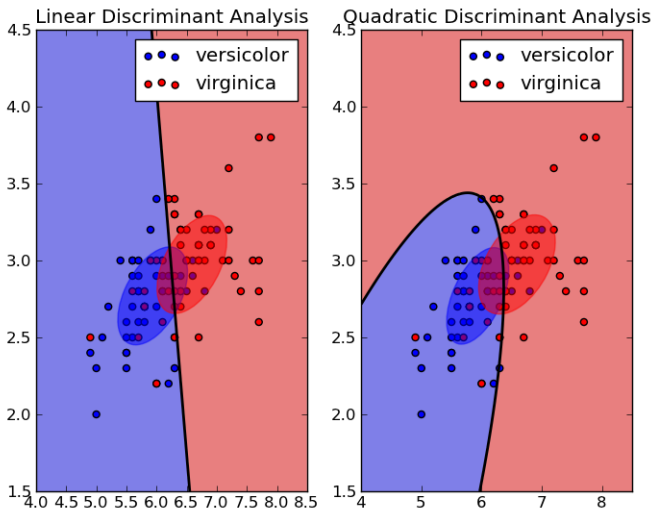

Figura 2.4: Fronteras definidas por LDA (derecha) y QDA (izquierda) en el conjunto clásico de las flores Iris [14]

2.2.3. Implementación utilizada

Al contrario que la mayoría del proyecto correspondiente a esta memoria, se ha realizado con Python debido a diversos fallos en los paquetes relacionados con selección de variables de la implementación de LDA de R [15, hace referencia a la manifestación del mismo error en la función de selección de variables empleada sobre una red neuronal].

Se ha utilizado el paquete sklearn [16] de Python con las funciones de discriminante correspondiente así como las funciones de estimación de matriz de confusión, tasa de acierto, especificidad, sensibilidad, tasa de falsos positivos y tasa de falsos negativos por validación cruzada. Con estos últimos se han podido representar las curvas ROC así como estimar la tasa de equierror. La función RFE ha servido para hacer una selección recursiva de variables [17].

Con respecto a los modelos teóricos desarrollados previamente, se han sustituido cada uno de los parámetros por sus estimadores muestrales $(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2)$.

La fuerte correlación entre variables en el conjunto de datos B, especialmente en el Corpus II, da lugar a problemas de alta variabilidad por multicolinealidad. En la mayoría de los casos, la función RFE ha servido para hacer una selección recursiva de variables [17] y con ello eliminar el problema. En otros casos en los que el clasificador resultante era de resultados inestables, se han eliminado paso a paso previamente a la aplicación de la función RFE aquellas variables cuyo factor de inflación de la varianza (VIF) fuese superior a 5 [18].

El factor de inflación de la varianza (VIF) nos sirve para conocer la influencia de una variable en la multicolinealidad de un conjunto de estas. Si contamos con una colección de variables $(X_1, ..., X_k)$ cuyos parametros en el ajuste toman valores $(\beta_1, ..., \beta_k)$ y llamamos R_j^2 al coeficiente de determinación de la regresión con la variable a explicar X_j y el resto de variables como regresores, podemos definir el VIF de una de las variables como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

2.3. Árboles de decisión

Un árbol de decisión [19] es el resultado de realizar una secuencia ordenada de disyunciones de forma que la disyuntiva hecha en cada paso dependerá de la respuesta dada a las anteriores [20]. La secuencia concluye en una predicción de la clase.

El punto de inicio se denomina 'nodo raíz' y cada nodo sucesor puede ser terminal o no-terminal. Un nodo no-terminal es siempre el nodo padre de dos nodos hijos, que son una división binaria determinada por una condición booleana aplicada sobre el valor de una variable única. Un nodo terminal es todo aquel que no se divide y se le asigna una etiqueta de clase. Al conjunto de todos los nodos terminales de un árbol se les llama partición de datos.

En cada nodo el algoritmo de crecimiento del árbol debe decidir qué variable es la más adecuada para ser la que se opere en ese nodo. Una función de impureza del nodo sirve como criterio a partir del cual podemos definir la mejor división de entre todas las variables en el mismo. En un conjunto de datos con 2 clases $(k = 0, 1)$ llamamos $P(k|\tau)$ a la estimación de probabilidad condicionada de que la observación sea de la clase k teniendo en cuenta que estamos en el nodo τ. La función de impureza del nodo será:

$$i(\tau) = \varphi(P(0|\tau), P(1|\tau))$$

Las dos funciones de impureza más usadas son, si establecemos que $p = P(1|\tau)$:

- Función de Entropía (definida con logaritmos en base dos):

$$i(\tau) = -\sum_{k=0}^{1} P(k|\tau) \cdot \log_2 P(k|\tau) = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$$

- Indice de Gini:

$$i(\tau) = \sum_{k \neq k'} P(k|\tau) \cdot P(k'|\tau) = 2p \cdot (1-p)$$

La Figura 2.5 muestra como el comportamiento esperado de las dos funciones de impureza es bastante similar.

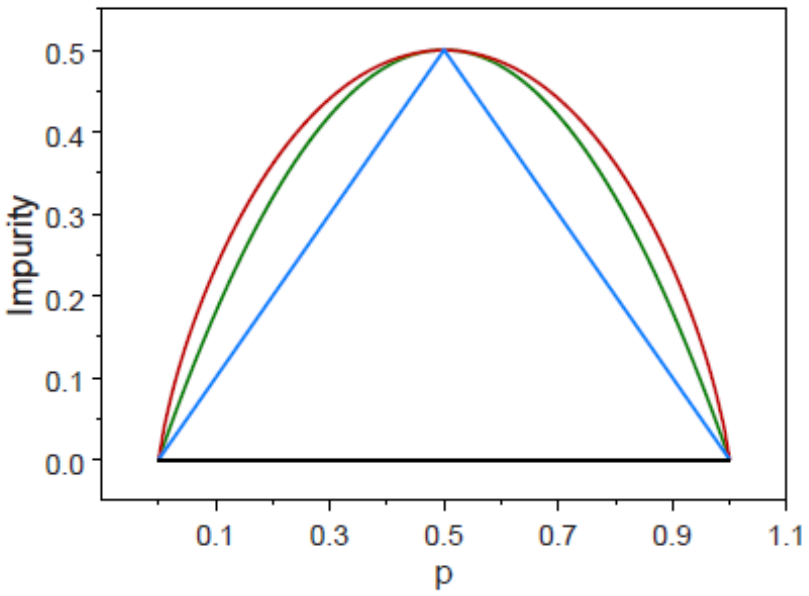

Figura 2.5: Comparación entre funciones de impureza. En rojo aparece la entropía escalada, en verde el índice de Gini y en azul la estimación de tasa de error.

Supongamos que en el nodo τ aplicamos una división según un valor c a una variable X_j de tal forma que quede una proporción de observaciones p_L hacia el nodo hijo τ_L y el resto p_R hacia el otro nodo hijo τ_R. Si la clase del conjunto de datos es dicotómica, se podrá establecer la Tabla 2.3:

	$Y=0$	$Y=1$	
τ_L: $X_j \leq c$	n_{11}	n_{12}	n_{1+}
τ_R: $X_j > c$	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

Tabla 2.3: División del conjunto de datos dicotómico según una condición sobre una variable X_j.

La función de entropía vendrá dada, si estimamos p_L como $\frac{n_{+1}}{n_{++}}$ y p_R como $\frac{n_{+2}}{n_{++}}$:

$$i(\tau) = -\frac{n_{+1}}{n_{++}} \cdot \log\left(\frac{n_{+1}}{n_{++}}\right) - \frac{n_{+2}}{n_{++}} \cdot \log\left(\frac{n_{+2}}{n_{++}}\right)$$

Los nodos hijo τ_R y τ_L tendrían asociadas las siguientes entropías si estimamos p_L como $\frac{n_{11}}{n_{1+}}$ y p_R como $\frac{n_{12}}{n_{1+}}$ para $X_j \leq c$ y p_L como $\frac{n_{21}}{n_{2+}}$ y p_R como $\frac{n_{22}}{n_{2+}}$ en caso contrario:

$$i(\tau_L) = -\frac{n_{11}}{n_{1+}} \cdot \log\left(\frac{n_{11}}{n_{1+}}\right) - \frac{n_{12}}{n_{1+}} \cdot \log\left(\frac{n_{12}}{n_{1+}}\right) \qquad i(\tau_R) = -\frac{n_{21}}{n_{2+}} \cdot \log\left(\frac{n_{21}}{n_{2+}}\right) - \frac{n_{22}}{n_{2+}} \cdot \log\left(\frac{n_{22}}{n_{2+}}\right)$$

Definimos con esto la ganancia de información para la división en el valor c de la variable X_j para el nodo τ. La mejor división para la variable será aquella que maximice este valor.

$$\Delta i(s,\tau) = i(\tau) - p_L \cdot i(\tau_L) - p_R \cdot i(\tau_R)$$

La aplicación en cada nodo de la ganancia de información da lugar al particionamiento recursivo según el cual se divide cada nodo del árbol hasta que no pueda dividirse más. En los nodos terminales la clase asignada será la mayoritaria. Sin embargo, permitir el crecimiento máximo al árbol de decisión lo sobredimensiona y da lugar a sobreajuste. Por ello se hace necesaria la poda de los árboles.

La poda de un árbol [6] permite la creación de un subárbol cuya tasa de error asociada sea inferior, aunque el error aparente sea más elevado. La tasa de error se estimará mediante validación cruzada, Leave-One-Out, conjunto test independiente,...

Existen dos tipo de poda: la prepoda, que busca impedir el crecimiento de una rama cuando la información de ésta no es suficientemente fiable y la postpoda que tras la creación del árbol completo elimina los subárboles cuya información asociada sea poco fiable. La poda más utilizada es la postpoda, ya que a cambio de un coste computacional mayor logra mejores resultados en la mayoría de los problemas. Las dos operaciones de postpoda son (Figura 2.6):

- Reemplazar subárbol (subtree replacement): consiste en reemplazar nodos internos por nodos hoja. Se comienza la búsqueda de candidatos por los nodos de mayor profundidad.
- Elevar subárbol (subtree raising): reemplazar un nodo padre por un subárbol redistribuyendose los individuos. Se aplica unicamente a las ramas más pobladas.

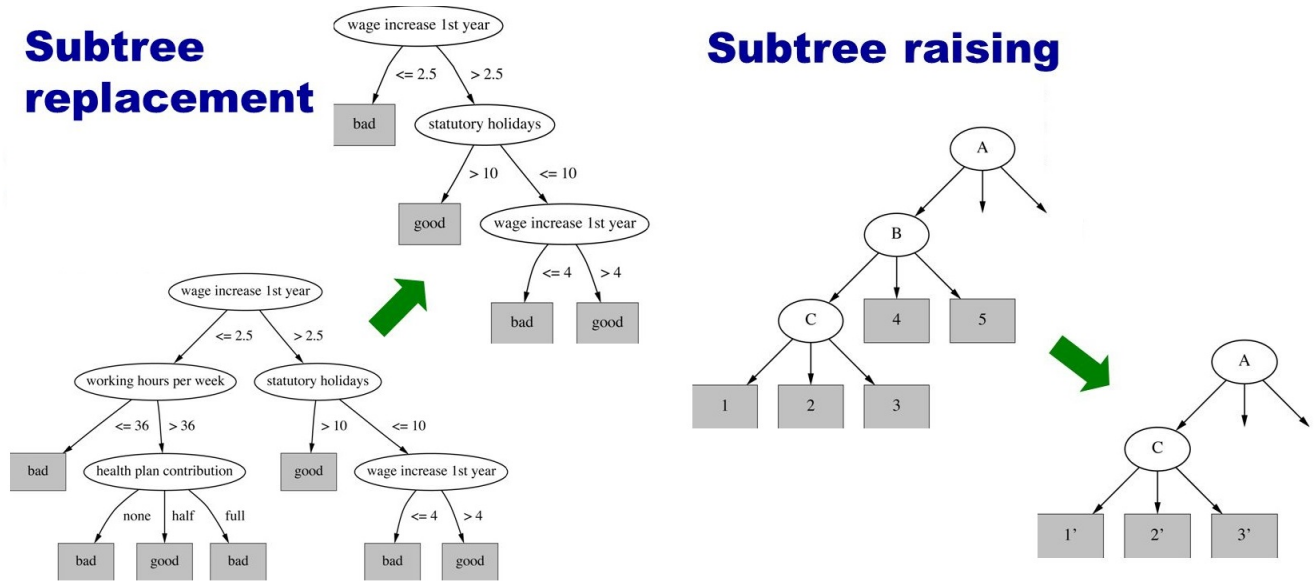

Figura 2.6: Operaciones básicas de la poda (postpoda) en árboles de decisión: reemplazar árbol (izquierda) y elevar subárbol (derecha) [21].

2.3.0.1. Implementación utilizada

Se ha desarrollado en R con ayuda del paquete auxiliar de aprendizaje Caret [22] y la implementación de árboles C4.5 con la interfaz de Weka para R (paquete RWeka [23]). Estos árboles trabajan con atributos continuos y controlan que el número de observaciones en los nodos terminales sea menor que un determinado número mediante un parámetro M. Realizan poda pesimista controlada por un parametro C con las dos operaciones de postpoda.

La salida del árbol se ha establecido que sea probabilística. Se ha empleado caret para realizar validación cruzada (10 particiones) y estimar la tasa de acierto, las matrices de confusión, tasa de falsos positivos y tasa de falsos negativos. Con estas últimas se ha podido representar con Python las curvas ROC medias y definir la tasa de equierror asociada al clasificador.

Mediante validación cruzada se han optimizado los valores de los parámetros C y M. Se ha representado el árbol de decisión a partir del lenguaje Dot y la herramienta Graphviz [24].

2.3.1. Bagging: Random Forest

Random Forest [25] es una técnica de análisis supervisado y de regresión fundamentada en la combinación de varios árboles de decisión [20]. Los orígenes de esta técnica datan de finales del año 1995 cuando su autor Tim Kam Ho propuso el fundamento teórico subyacente a estos modelos.

La limitación principal de los árboles de decisión suele ser su alta variabilidad debido a su tendencia a sobreajustar el conjunto de entrenamiento [7]. Una solución posible a esto es aplicar bagging (bootstrap aggregating) en el que se combinan muchos clasificadores cuyo error aparente es bajo.

El fundamento de Random Forest es el bagging así como un procedimiento de selección de variables aleatorio para inducir los árboles. Para desarrollar cada uno de los árboles se selecciona un subconjunto de variables cuyo tamaño suele ser cercano a la raíz cuadrada del número de características.

La predicción de la clase $\hat{y}$ se resuelve mediante lo que sería un voto mayoritario, es decir, la clase más veces predicha por cada uno de los árboles de decisión base subyacentes al procedimiento Random Forest será el valor que tomará $\hat{y}$. La estructura general del modelo la podemos observar en la Figura 2.7.

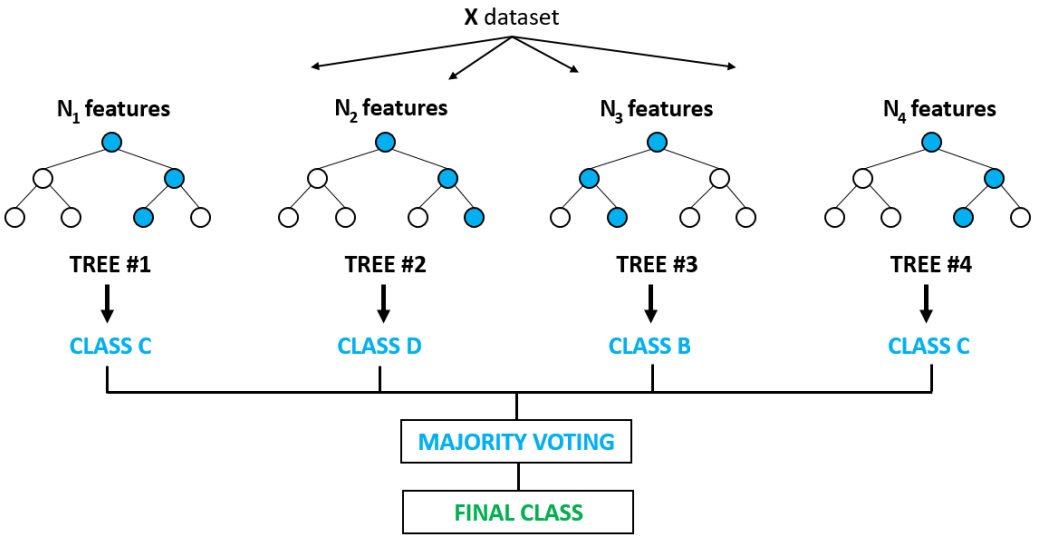

Figura 2.7: Estructura de los modelos Random Forest [26]

Una de las principales ventajas de este clasificador es la posibilidad de conocer la importancia relativa de cada una de las variables de entrada al modelo. Esta importancia relativa se calcula a partir del error 'Out-Of-Bag' (OOB) en el que promedia el error de predicción de una muestra x_i que no ha formado parte de la muestra bootstrap de entrenamiento en esa iteración.

Las aplicaciones de Random Forest son muchas, aumentando estas prácticamente de forma diaria. Muchos artículos científicos han empleado esta metodología en diferentes ámbitos, como pueden ser los sistemas de clasificación de imágenes. Wikipedia, por ejemplo, los emplea como criterio de calidad para sus páginas. Existen también desarrollos de Random Forest para análisis no-supervisado como medida de disimilaridad entre los datos[27].

2.3.1.1. Implementación utilizada

Se ha desarrollado en R con ayuda del paquete auxiliar de aprendizaje Caret [22] y su principal implementación de Random Forest: el paquete randomForest [28]. Se ha empleado la función rfeControl de Caret para optimizar el número de variables del modelo así cómo las funciones nativas de este paquete para realizar validación cruzada y estimar la tasa de acierto, las matrices de confusión estimadas, tasa de falsos positivos y tasa de falsos negativos. Mediante validación cruzada se ha optimizado también el parámetro del modelo que controla el número de variables aleatorias a emplear por partición. Con las tasas de falsos positivos y falsos negativos calculadas en R se ha calculado la tasa de equierror y, con ayuda del lenguaje de programación Python, se han representado las curvas ROC correspondientes.

2.3.2. Boosting: Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost)[19] es una técnica de análisis supervisado fundamentada en la combinación ('Boosting') de árboles de decisión[7]. Los estudios de Yoav Freund y Robert Schapire publicados en el año 2003 dieron lugar a estos modelos. Con ellos pudieron ganar el premio Gözel de reconocimiento en avances en ciencias de la computación.

Llamamos Boosting [20] a la técnica de regresión y análisis supervisado fundamentada en la combinación y remuestreo de varios árboles de decisión simples de poca variabilidad pero de error aparente muy elevado (clasificadores débiles).

La predicción de la clase $\hat{y}$ se resuelve mediante la ponderación de las salidas de los clasificadores débiles que sirven como base al clasificador AdaBoost. La estructura general del modelo la podemos observar en la Figura 2.8.

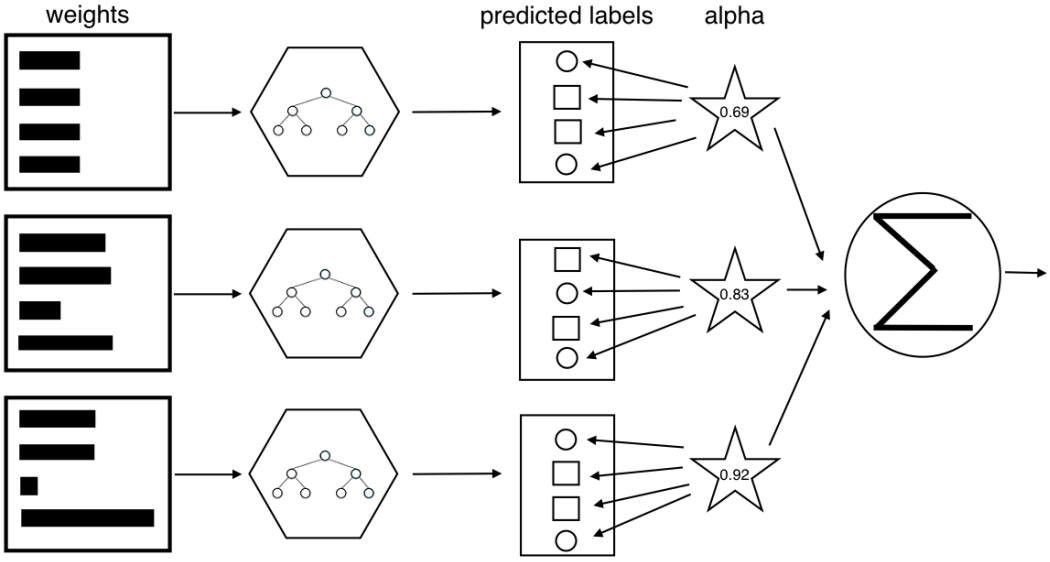

Figura 2.8: Estructura de los modelos AdaBoost [29]

Es un algoritmo creado para poderse enfrentar a la casuística propia de muchos conjuntos de datos de Big Data: el número de variables o atributos es muy elevado en comparación con el número de observaciones. Al emplearse múltiples clasificadores débiles, el clasificador intrínsecamente seleccionará solo aquellas variables que realmente le den capacidad discriminatoria e ignorará a las que no tengan este efecto.

En estos modelos será de relevancia ajustar los parámetros de los árboles básicos subyacentes (profundidad y número) así como un metaparámetro α relacionado con el aprendizaje de los propios parámetros del modelo. Este pondera las salidas de los árboles de tal forma que se dé más peso a las observaciones mal clasificadas. En función del algoritmo empleado para la optimización, α tomará los siguientes valores a partir de la tasa de error (e) y el número de clases (k):

- Breiman:
$$\alpha = \frac{1}{2}\log\left(\frac{1-e}{e}\right)$$
- Freund:
$$\alpha = \log\left(\frac{1-e}{e}\right)$$
- Zhu:
$$\alpha = \log\left(\frac{1-e}{e}\right) + \log(k-1)$$

Existen múltiples aplicaciones para los modelos AdaBoost. El origen de éstos fue el primer algoritmo de reconocimiento de objetos en tiempo real jamás creado, el algoritmo de Viola-Jones [30]. El proyecto inicial buscaba crear un sistema de reconocimiento facial fiable.

2.3.2.1. Implementación utilizada

Se ha desarrollado en R con ayuda del paquete auxiliar de aprendizaje Caret [22] y uno de los paquetes de Boosting más empleados: adabag [31]. Mediante las funciones auxiliares de Caret de validación cruzada (10 particiones) se han estimado la tasa de acierto, las matrices de confusión, tasa de falsos positivos y tasa de falsos negativos. Con estas últimas se ha calculado la tasa de equierror y haciendo uso de Python se han representado las curvas ROC correspondientes.

Mediante validación cruzada se han determinado los valores óptimos para los parámetros del modelo relacionados con la profundidad y número de árboles básicos así como el procedimiento óptimo de aprendizaje mediante el ajuste de α. La salida básica de adabag devuelve la importancia relativa de las variables. Se han hecho representaciones con esto para cada uno de los modelos.

2.3.3. Boosting: Bayesian Additive Regression Trees (BART)

Los modelos BART (Bayesian Additive Regression Trees) son relativamente recientes, datándose los primeros usos de estos cerca del año 2008 [32]. Se trata de modelos robustos muy prometedores que suman secuencialmente la contribución de varios clasificadores débiles mediante Boosting [33].

Se les apoda cómo bayesianos porque en este aprendizaje secuencial previamente mencionado se usa la noción de a priori y de verosimilitud para calcular probabilidades a posteriori. Esto ofrece varias ventajas, cómo que la formulación bayesiana intrínsecamente realiza regularización y evita así árboles sobreajustados.

Debido a la complejidad de las a posteriori de los BART, son necesarias técnicas de MCMC (Monte Carlo Markov Chains) para simularlas. Los algoritmos de simulación empleados en la mayoría de implementaciones de BART son Metropolis-Hasting [25] y el muestreo de Gibbs [25].

La poda de los árboles que conforman el modelo se controla mediante el hiperparámetro base α y el hiperparámetro exponente β. Estos controlan el que un nodo de los árboles sea terminal o no.

Al fundamentarse en la combinación de múltiples árboles en los que se seleccionan variables útiles para la clasificación, se puede emplear el porcentaje de veces que se ha utilizado esa variable para inducir el árbol como medida de la importancia relativa de la misma. La Figura 2.9 muestra un ejemplo de modelo BART que refleja claramente la naturaleza bayesiana aditiva intrínseca a la técnica.

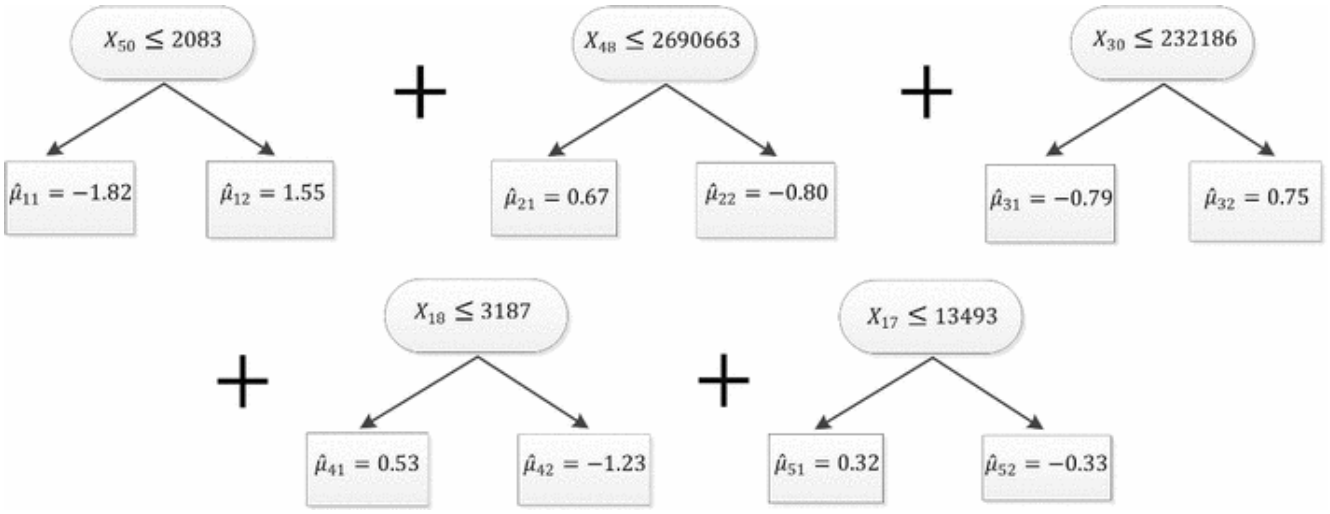

Figura 2.9: Estructura de un modelo BART en el conjunto de datos 'cáncer de próstata' [34]

Una de las curiosidades de este tipo de metodología es cómo el número de árboles a generar no se decide a priori, sino que se computa de forma dinámica. Sin embargo, en la mayoría de problemas no suele sobrepasar los 200 árboles [33]. La desventaja de esta metodología es el elevado tiempo de cómputo.

Al ser modelos de reciente creación se están poco a poco introduciendo en diversos ámbitos y logrando resultados altamente satisfactorios. Alguno de los campos en los que ha logrado mejores resultados en modelado de riesgo en créditos [35] y en análisis en tiempo real de ADN vírico [35].

2.3.3.1. Implementación utilizada

Se ha desarrollado en R con ayuda del paquete auxiliar de aprendizaje Caret [22] y la única implementación BART de R con el paquete bartMachine [32]. Se ha empleado caret para realizar validación cruzada y calcular la tasa de acierto y las matrices de confusión estimadas. También se ha usado para estimar la tasa de equierror y tasas de falsos positivos y falsos negativos que se han empleado en Python para la representación de curvas ROC.

Mediante validación cruzada se han seleccionado los parámetros óptimos: el número total de árboles que inducir y los hiperparámetros de poda (α y β).

La función investigateVarImportance de bartMachine ha servido para estimar la importancia relativa de las variables con la frecuencia de selección de cada una de ellas para la inducción de los árboles básicos de decisión.

Capítulo 3

Resultados

En esta sección se valorará el rendimiento obtenido por cada uno de los clasificadores, entendiendo el rendimiento como nivel de acierto obtenido con el modelo en cuestión correspondiente a uno de los Corpus. Será relevante comparar los resultados de los clasificadores con el rendimiento obtenido por "Smart Eyes", el sistema en funcionamiento en Tradema, y con un sistema de clasificación de referencia. Esto será una clasificación sencilla con un umbral, tal y como aparece descrito en el Apartado 3.2.

3.1. Sistema "Smart Eyes"

El número de tableros que clasifica mal (tanto falsos positivos como falsos negativos) es en principio desconocido. Por lo tanto, se trata de un sistema en el que no existe ningún tipo de medida de rendimiento de forma inmediata. Aunque detecta muchos defectos, crea también muchos avisos que son falsos positivos.

Todos los resultados de rendimiento que obtendremos en este proyecto de este sistema están muy sesgados: son datos obtenidos únicamente por las observaciones del Corpus I. Este Corpus, tal y como se explica en el Apartado **??**, se compone solo de imágenes que el sistema ha etiquetado como defectuosas. Definimos un tipo de defecto detectado a mayores que se da en la salida de "Smart Eyes":

- **Defecto detectado por fallo lumínico**: deficiencia del sistema "Smart Eyes" relacionada con el filtrado de blancos del sistema en imágenes con problemas de heterogeneidad de luz. Se da de forma distinta en imágenes sin defectos que en las que los tienen:

    - En imágenes sin defectos, "Smart Eyes" detecta defectos por la diferencia de luz donde realmente no los hay como tal. Las Figuras 3.1 y 3.2 muestran un ejemplo de imagen sin defectos en la que el sistema filtrando ha cometido un error por fallo lumínico. Las Figuras 3.3 y 3.4 muestran otro caso más extremo donde la excesiva iluminación ha dado lugar a un incorrecto etiquetado de la imagen como defectuosa.

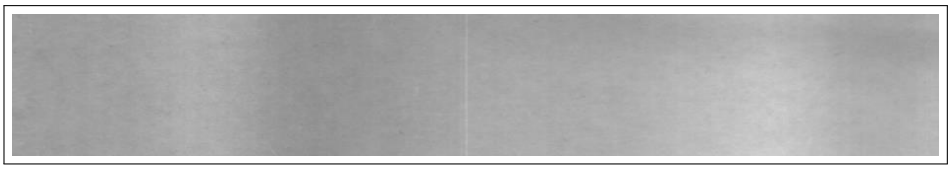

Figura 3.1: Imagen sin defectos con problemas en la homogeneidad de la iluminación.

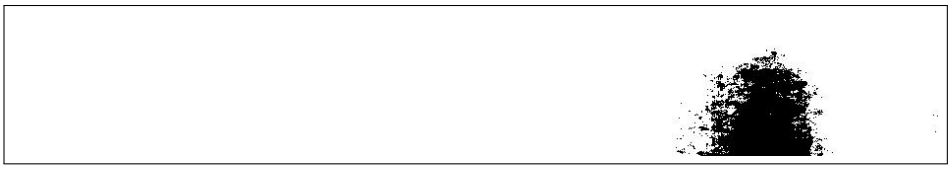

Figura 3.2: Imagen filtrada. La zona con más brillo se detecta erróneamente como defectuosa.

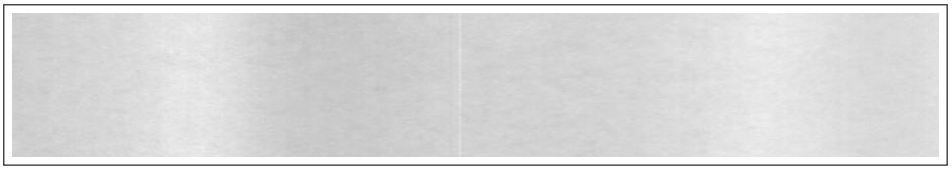

Figura 3.3: Imagen sin defectos con más problemas de homogeneidad de luz.

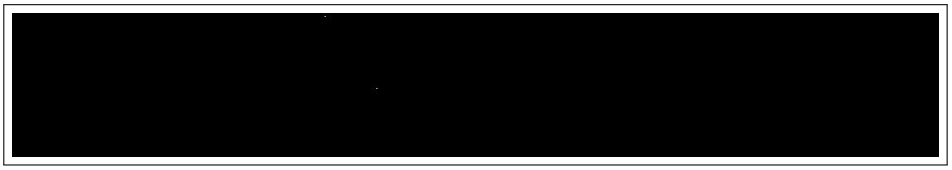

Figura 3.4: Imagen filtrada. El sistema tiene aquí un comportamiento errático.

- En los casos de imágenes con defectos, "Smart Eyes" la identifica correctamente como defectuosa. Sin embargo las áreas de error se marcan de forma errónea por lo que es en cierto modo una detección fortuita del defecto, tal y como podemos observar en las Figuras 3.5 y 3.6

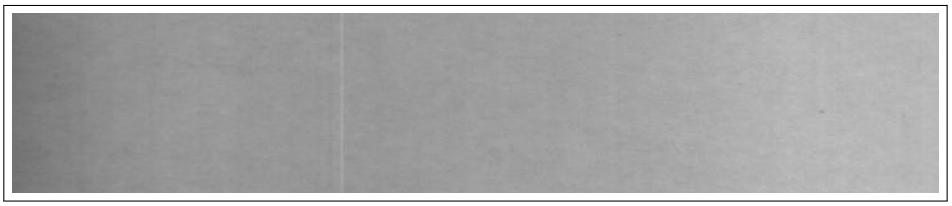

Figura 3.5: Imagen con problemas de iluminación que contiene un defecto negro de pequeño tamaño.

Figura 3.6: Imagen filtrada. El sistema detecta defecto, pero no ha marcado como defecto el auténtico defecto.

Con la definición de este error, podemos determinar la matriz de confusión de este sistema. La matriz de confusión aparece en la Tabla 3.1 así como su versión simplificada en la Tabla 3.2. Es relevante recordar que no tenemos acceso a las imágenes que ha detectado "Smart Eyes" como carentes de defectos, lo que da lugar a que ambas matrices de confusión tengan una fila sin datos.

Matriz de confusión de "Smart Eyes"					
		Defecto real en la imagen			
		Sin defectos	Negro	Blanco	Negro+Blanco
Defec. detectado	Sin defectos	-	-	-	-
	Negro	0	5475	0	2
	Blanco	61	1	520	1
	Negro+Blanco	0	0	0	451
	Fallo lumínico	648	13	4	5

Tabla 3.1: Matriz de confusión del sistema "Smart Eyes".

Matriz de confusión simplificada de "Smart Eyes"			
		Clase	
		Sin defecto	Con defecto
Etiq.	Sin defecto	-	-
	Con defecto	709	6458

Tabla 3.2: Matriz de confusión simplificada (clase de la imagen y clasificación asignada) del sistema "Smart Eyes".

Podemos hacer algunas valoraciones con respecto al filtrado hecho por "Smart Eyes" a la vista de los resultados obtenidos:

- El filtrado de negro funciona bastante bien. Aún así detecta zonas especialmente oscuras y a veces zonas defectuosas más claras no son correctamente marcadas. Se genera ruido en el filtrado habitualmente. En las Figuras 3.7 y 3.8 podemos ver un ejemplo de estas dos particularidades mentadas.

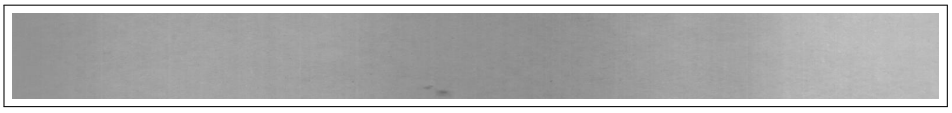

Figura 3.7: Imagen del Corpus I con defectos negros de distinta oscuridad.

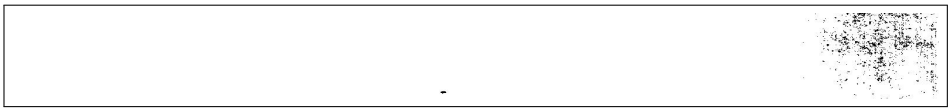

Figura 3.8: Imagen filtrada. Observamos ruido a la derecha y zonas no detectadas como defectuosas.

- El filtrado de blancos, aún siendo más complejo que el de negros, tiene muchos problemas con las imágenes en las que la luz no es homogénea. Esto da lugar al defecto por fallo lumínico previamente mencionado.

Teniendo en cuenta la matriz de confusión podemos hacer una estimación parcial altamente sesgada de la tasa de error. Este sesgo se deberá a la ausencia de datos de las imágenes sin defectos.

Tasa parcial de acierto estimada de "Smart Eyes"		
	Tasa de acierto	Intervalo de confianza del 95 %
Smart Eyes	90.107 %	(89.416, 90.798)

Tabla 3.3: Estimación altamente sesgada de la tasa de acierto para el sistema en funcionamiento en la fábrica -"Smart Eyes"-.

3.2. Sistema de referencia

Siguiendo las pautas expuestas en la sección 2.1, se han evaluado las curvas ROC y la tasa de equierror obtenidas empleando independientemente el área de error relativa hallada en el filtrado de negros y en el de blancos y la obtenida empleando la suma de ambas (área de error relativa en blanco + área de error relativa en negro). La definición de umbrales y de resultados obtenidos se hará por separado para el Corpus I y el II.

Posteriormente se han definido modelos con un umbral para el estudio más concreto sobre los que definir la tasa de acierto, especificidad y sensibilidad.

3.2.1. Corpus I

En el caso del Corpus I, se obtienen resultados básicos altamente satisfactorios, tal y como podemos observar en las curvas ROC (área de error relativa en blanco y área de error relativa en negro por separado Figura 3.9; con la suma de ambas: Figura 3.10) así como en los valores de la tasa de equierror y AUC (Tabla 3.4). Esto se debe primordialmente al excelente funcionamiento del filtrado en imágenes cuya resolución es muy alta.

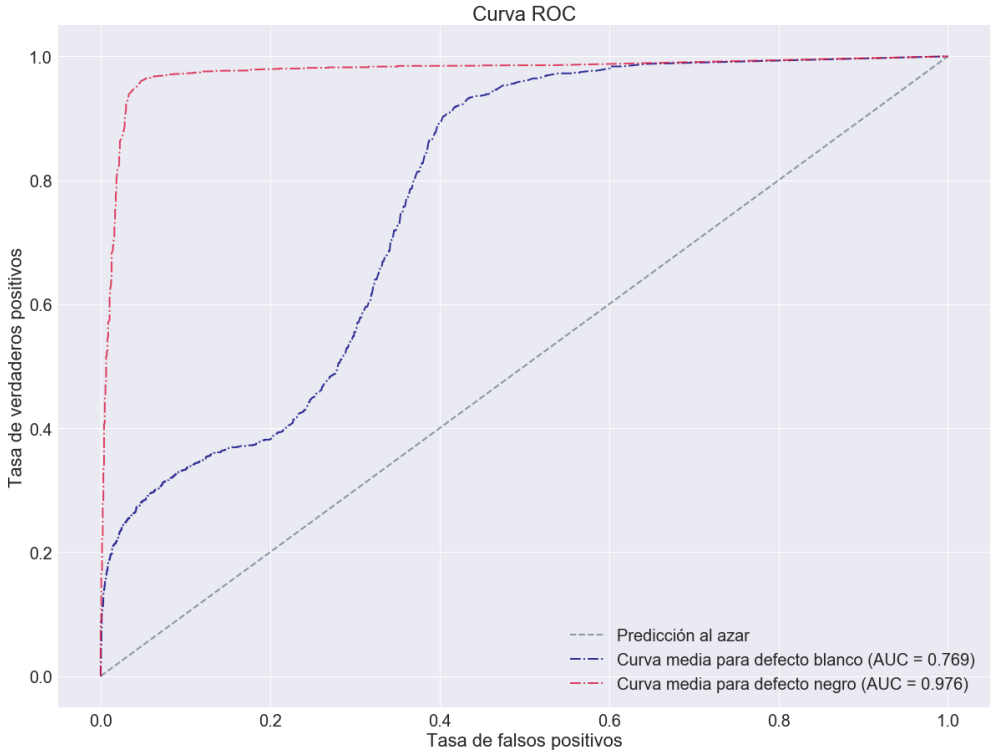

Figura 3.9: Curvas ROC del sistema de referencia empleando por separado área de error relativa en blanco (azul) y área de error relativa en negro (rojo) del Corpus I.

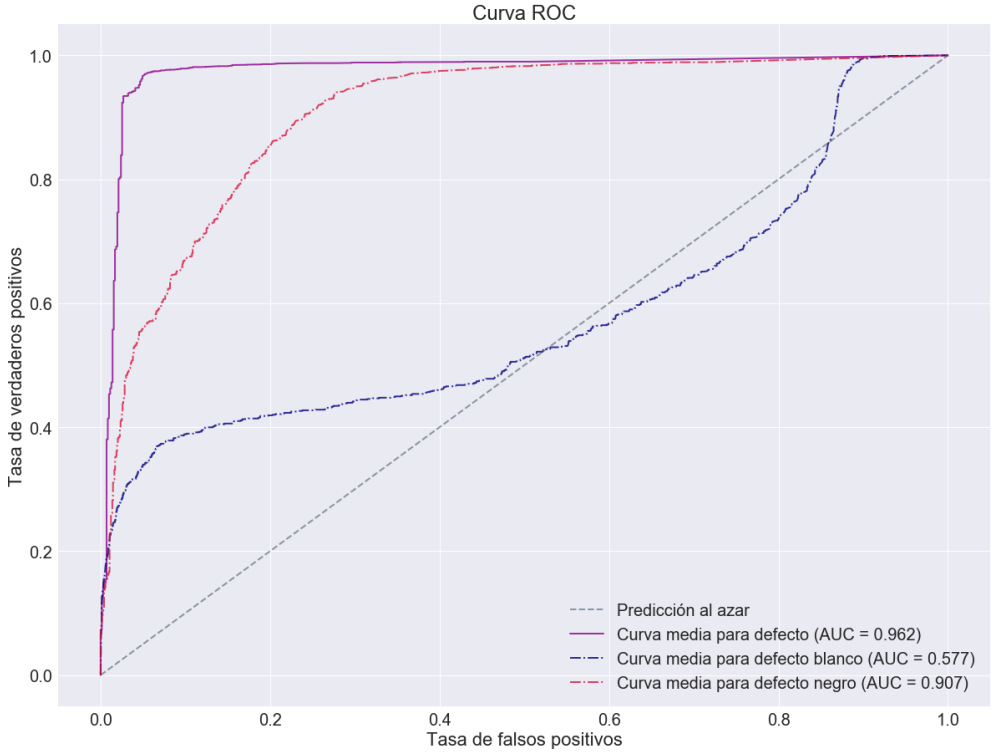

Figura 3.10: Curva ROC del sistema de referencia empleando la suma de las áreas de error relativas en negro y blanco para la clasificación con defecto - sin defecto del Corpus I.

Resultados del sistema de referencia en el Corpus I				
		Tasa de equierror (EER)	Area Under Curve (AUC)	
Suma de áreas de error		0.046	0.962	Negro: 0.907 Blanco: 0.577
Áreas de error por separado	Negro	0.044	0.976	
	Blanco	0.333	0.769	

Tabla 3.4: Estimación de la tasa de equierror y el área bajo la curva ROC del sistema de referencia con el Corpus I. En el caso de usarse la suma de áreas de error se ha calculado el AUC para cada clase de error.

En el caso del uso de la suma de las áreas de error relativas, los valores de la AUC y de la EER son excelentes. La posterior comparativa entre él área bajo la curva ROC de las clases con defecto negro-sin defecto negro y la correspondiente a con defecto blanco-sin defecto blanco pone en evidencia el funcionamiento deficiente del sistema con los defectos blancos.

Usando las dos áreas de error relativa por separado se obtiene un sistema de características similares al anterior. La discriminación de defectos blancos tan solo por el área de error relativa es compleja, llegando a mal clasificarse cerca del 30 % de las observaciones al igualarse las tasas de falsos positivos y falsos negativos.

Los modelos de umbral ajustado según la tasa de equierror seleccionados obtienen resultados básicos altamente satisfactorios, tal y como podemos observar en las tablas de resultados (Tablas 3.5 y 3.6).

Sistema de referencia (1 umbral) -Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	676	271	947
	Con defecto	33	6187	6220
		709	6458	7167

- Tasa de acierto: 95.75 %
  Intervalo de confianza del 95 %: (95.28, 96.21)
  - Sensibilidad: 95.80 %
  - Especificidad: 95.34 %

Tabla 3.5: Matriz de confusión y tasa de acierto con el uso de 1 umbral (área relativa de error en negro + área relativa de error en blanco) en Corpus I.

Sistema de referencia (2 umbrales) -Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	684	286	970
	Con defecto	25	6172	6197
		709	6458	7167

- Tasa de acierto: 95.66 %
  Intervalo de confianza del 95 %: (95.19, 96.13)
  - Sensibilidad: 95.57 %
  - Especificidad: 96.47 %

Tabla 3.6: Matriz de confusión y tasa de acierto con el uso de 2 umbrales (uno para área relativa de error en negro y otro para la de blanco) en Corpus I.

Las matrices de confusión reflejan cómo con dos umbrales el número de falsos positivos es menor que con uno, mientras que la tasa de falsos negativos es inferior con un umbral. Las tasas de acierto nos indican que el uso de dos umbrales da lugar a resultados globales ligeramente mejores. La especificidad y sensibilidad de los modelos son elevadas y equilibradas

3.2.2. Corpus II

Con el Corpus II los resultados que obtenemos con el sistema de referencia son menos satisfactorios, tal y como vemos en las curvas ROC (con dos umbrales: Figura 3.11; con un umbral: Figura 3.12) así como en los valores de la tasa de equierror y AUC (Tabla 3.7). Al haber sido la fuente de estas imágenes un vídeo, la resolución de los fotogramas extraídos es menor que la del Corpus I.

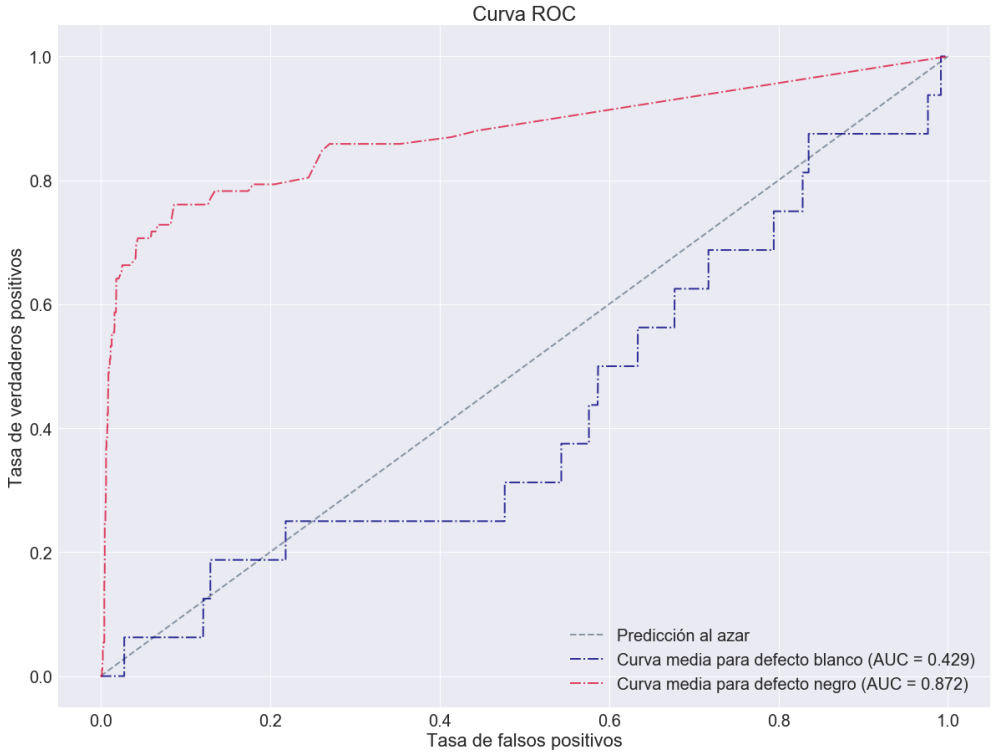

Figura 3.11: Curvas ROC del sistema de referencia empleando por separado área de error relativa en blanco (azul) y área de error relativa en negro (rojo) del Corpus II.

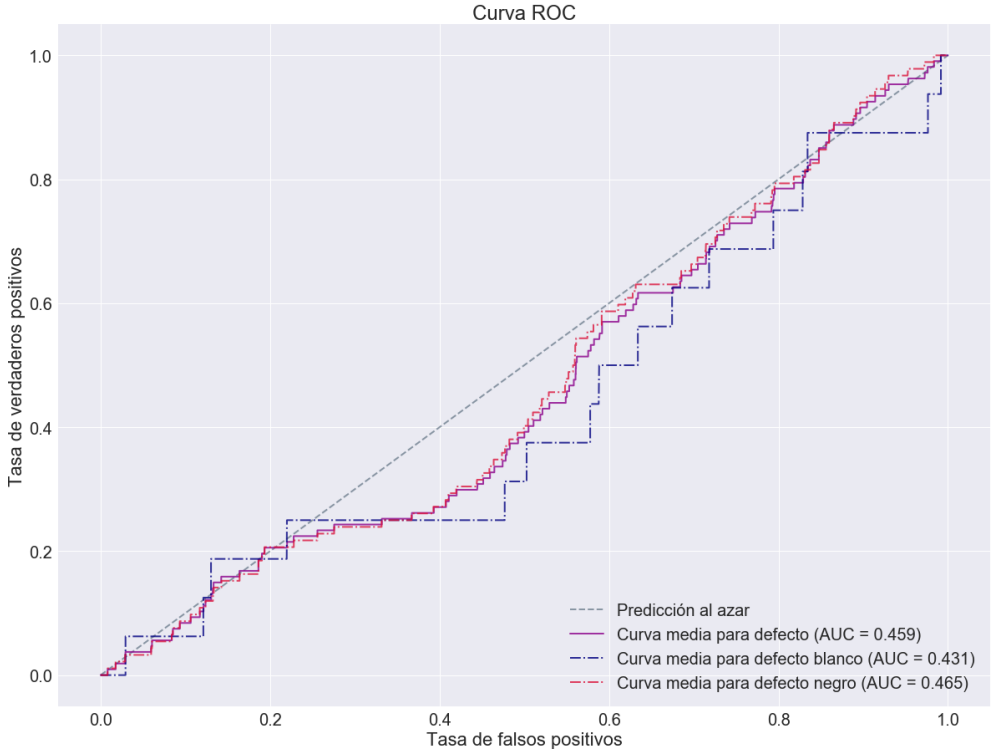

Figura 3.12: Curva ROC del sistema de referencia empleando la suma de las áreas de error relativas en negro y blanco para la clasificación con defecto - sin defecto del Corpus II.

Resultados del sistema de referencia en el Corpus II				
		Tasa de equierror (EER)	Area Under Curve (AUC)	
Suma de áreas de error		0.551	0.459	Negro: 0.465 Blanco: 0.431
Áreas de error por separado	Negro	0.205	0.872	
	Blanco	0.576	0.429	

Tabla 3.7: Estimación de la tasa de equierror y el área bajo la curva ROC del sistema de referencia con el Corpus II.

Empleando la suma de las áreas de error los valores del AUC y de la EER son bastante pobres, dando lugar a un sistema que clasificaría mal más del 50 % de las observaciones en el punto de la tasa de equierror. El uso por separado de las áreas de error relativas destaca de nuevo cómo la clasificación de los defectos blancos es mucho más complicada. La clasificación para los defectos negros es mejor atendiendo tanto al AUC como a la EER.

Los modelos de umbral ajustado seleccionados mediante tasa de equierror obtienen resultados mediocres, tal y como podemos observar en las matrices de confusión y en la tasa de acierto estimada (Tablas 3.8 y 3.9).

Sistema de referencia (1 umbral) -Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	1748	54	1802
	Con defecto	2145	53	2198
		3893	107	4000

- Tasa de acierto: 44.95 %
  Intervalo de confianza del 95 %: (40.07, 49.82)
  - Sensibilidad: 49.53 %
  - Especificidad: 44.90 %

Tabla 3.8: Matriz de confusión y tasa de acierto con el uso de 1 umbral (área relativa de error en negro + área relativa de error en blanco) en Corpus II.

Sistema de referencia (2 umbrales) -Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	1997	97	2094
	Con defecto	1896	10	1906
		3893	107	4000

- Tasa de acierto: 50.17 %
  Intervalo de confianza del 95 %: (45.27, 55.07)
  - Sensibilidad: 9.34 %
  - Especificidad: 51.29 %

Tabla 3.9: Matriz de confusión y tasa de acierto con el uso de 2 umbrales (uno para área relativa de error en negro y otro para la de blanco) en Corpus II.

Los resultados parecen indicarnos que emplear un umbral da lugar a una clasificación más equilibrada. Los valores de la sensibilidad y especificidad son bajos, especialmente en el modelo de dos umbrales dónde la especificidad roza el 10 %.

Por lo tanto, mientras que en el Corpus I se buscará refinar la precisión de un sistema que ya funciona bien por umbrales, en el Corpus II aún se puede mejorar mucho.

3.3. Análisis Discriminante

Siguiendo las pautas de la sección 2.2 se ha valorado la precisión de varios modelos de discriminante lineal y cuadrático. Sobre el total de las variables se ha realizado una selección recursiva de variables tal y como se desarrollo en el apartado de la implementación usada.

3.3.1. Discriminante Lineal (LDA)

Corpus I

3.3.1.1. Conjunto de datos A

Modelo óptimo: 35 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorN y AreaErrorB.
- PCA: CompPrinOr1, CompPrinOr3, CompPrinOr4, CompPrinTratB1, CompPrinTratB2, CompPrinTratN1, CompPrinTratN2.
- DCT: imagen sin binarizar: 1, 2, 3, 4, 6, 7, 11, 12, 16, 20, 22, 27, 28, 29; imagen binarizada: 2, 3, 4, 6, 7, 11, 16, 20, 22, 28, 29.

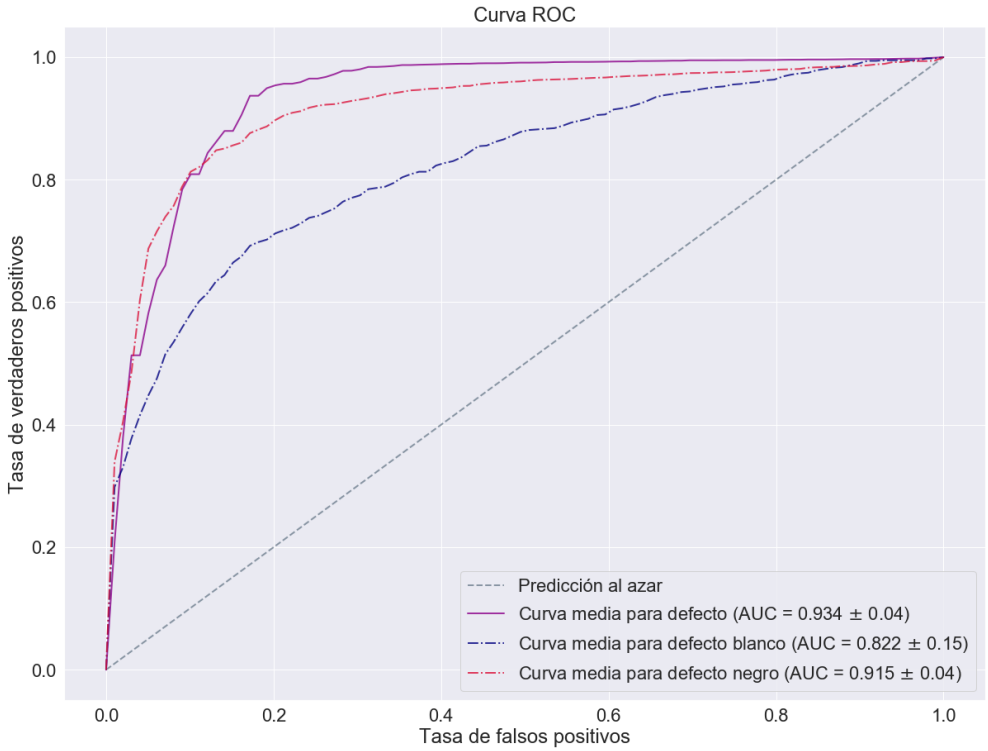

Figura 3.13: Curvas ROC de discriminante lineal empleando el conjunto de datos A del Corpus I.

Resultados de la discriminante lineal en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.103 ± 0.05	0.934 ± 0.04 {	Negro: 0.915 ± 0.04 Blanco: 0.822 ± 0.15

Tabla 3.10: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante lineal creado a partir del conjunto A de datos del Corpus I.

Las curvas ROC (Figura 3.13) así como las métricas relacionadas incluidas en la Tabla 3.10 nos indican como el sistema resultante de la discriminación lineal logra resultados algo peores que el sistema de referencia (mayor tasa de equierror y menor AUC). Sin embargo, la detección de defectos blancos parece ser mejor con este algoritmo.

Una vez ajustado el umbral probabilístico de clasificación, sean estimado diversas medidas de rendimiento que aparecen en la Tabla 3.10. La tasa de acierto es muy similar o ligeramente peor a la del sistema de referencia (Tabla 3.11). Mientras que el número de falsos negativos se ha reducido a la mitad, el número de falsos positivos ha crecido notablemente.

Sistema de discriminante lineal -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	483	111	594
	Con defecto	226	6347	6573
		709	6458	7167

- Tasa de acierto: 95.29 %.
  Intervalo de confianza del 95 %: (93.51, 97.09)
  - Sensibilidad: 98.28 %
  - Especificidad: 68.12 %

Tabla 3.11: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante lineal del Corpus I con el conjunto de datos A.

3.3.1.2. Conjunto de datos B

Modelo óptimo: 41 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorB y AreaErrorN.
- Estadísticos: MediaOr, VarianzaOr, KurtosisOr, MedianaOr, P10Or, P90Or, MediaTratadaB, VarianzaTratadaB, SkewnessTratadaB, KurtosisTratadaB, P10TratadaB, P90TratadaB, MediaTratadaN, VarianzaTratadaN, SkewnessTratadaN, KurtosisTratadaN, P10TratadaN P90TratadaN.
- DCT: imagen sin binarizar: 1, 2, 3, 4, 6, 11, 12, 15, 20, 22, 28, 29; imagen binarizada: 1, 2, 3, 4, 11, 15, 20, 22, 28, 29.

Resultados de la discriminante lineal en el Corpus I

	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.100 ± 0.04	0.944 ± 0.04 {	Negro: 0.931 ± 0.05 Blanco: 0.837 ± 0.14

Tabla 3.12: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante lineal creado a partir del conjunto B de datos del Corpus I.

Las curvas ROC asociadas al sistema de discriminante lineal (Figura 3.14) evidencian que la precisión resultante en lo que se refiere al defecto general es similar o algo peor que la del sistema de referencia. La detección de blancos de nuevo parece ser mejor. El valor de la tasa de equierror estimado es de un 10 % (Tabla 3.12).

En el modelo ajustado el resultado es muy similar al obtenido con el conjunto A: tasa de acierto muy similar o algo peor a la del sistema de referencia gracias a la disminución de falsos negativos y al agudo aumento de falsos positivos (Tabla 3.13). Este clasificador, con más variables que el óptimo del conjunto A, parece lograr una mayor tasa de acierto. Sin embargo, la especificidad es inferior.

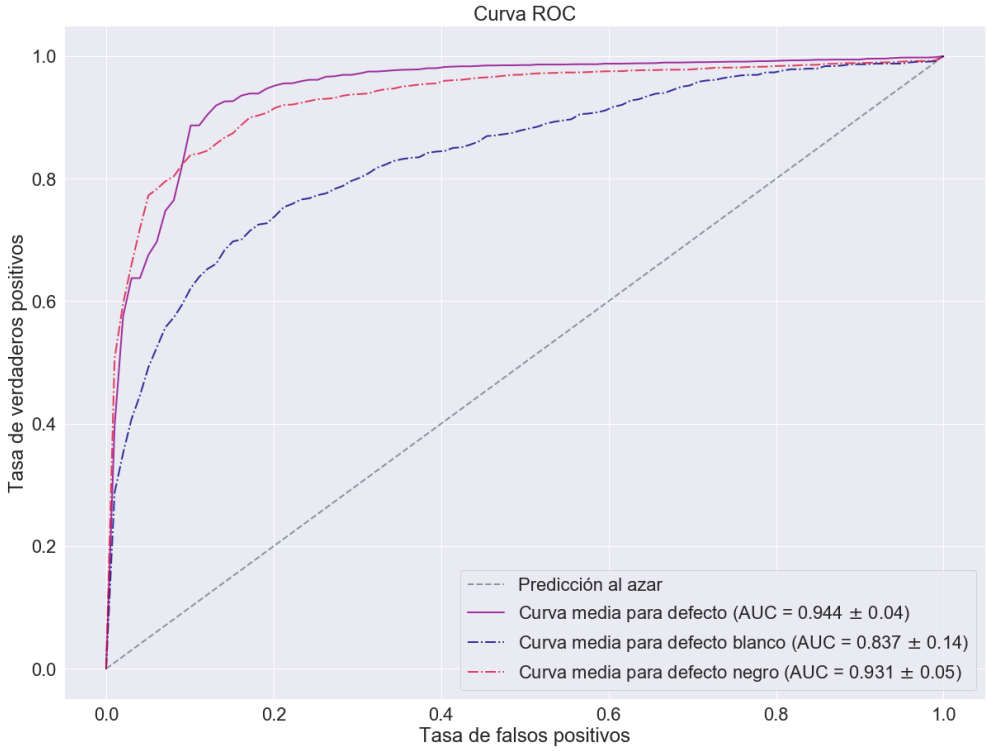

Figura 3.14: Curvas ROC de discriminante lineal empleando el conjunto de datos B del Corpus I.

Sistema de discriminante lineal -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	395	110	505
	Con defecto	314	6348	6573
		709	6458	7167

- Tasa de acierto: 94.08 %.
  Intervalo de confianza del 95 %: (92,09, 96.07)
  - Sensibilidad: 98.29 %
  - Especificidad: 55.71 %

Tabla 3.13: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante lineal del Corpus I con el conjunto de datos B.

Corpus II

3.3.1.3. Conjunto de datos A

Modelo óptimo: 35 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorN
- PCA: CompPrinTratB1.
- DCT: imagen sin binarizar: 1, 15, 16, 28; imagen binarizada: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30.

La Figura 3.15 con las curvas ROC y la Tabla 3.14 con medidas relacionadas a éstas indican que la discriminación lineal con el conjunto de datos B del Corpus II da lugar a una clasificación ligeramente mejor que la del sistema de referencia y la asignación al azar.

El modelo con el umbral de clasificación asignado es desequilibrado: la mejora sobresaliente la tasa de acierto con respecto al funcionamiento del sistema de referencia (Tabla 3.15) no es significativa cuando tenemos en cuenta que la sensibilidad es inferior al 10 %. El número de tableros defectuosos detectados correctamente es muy bajo.

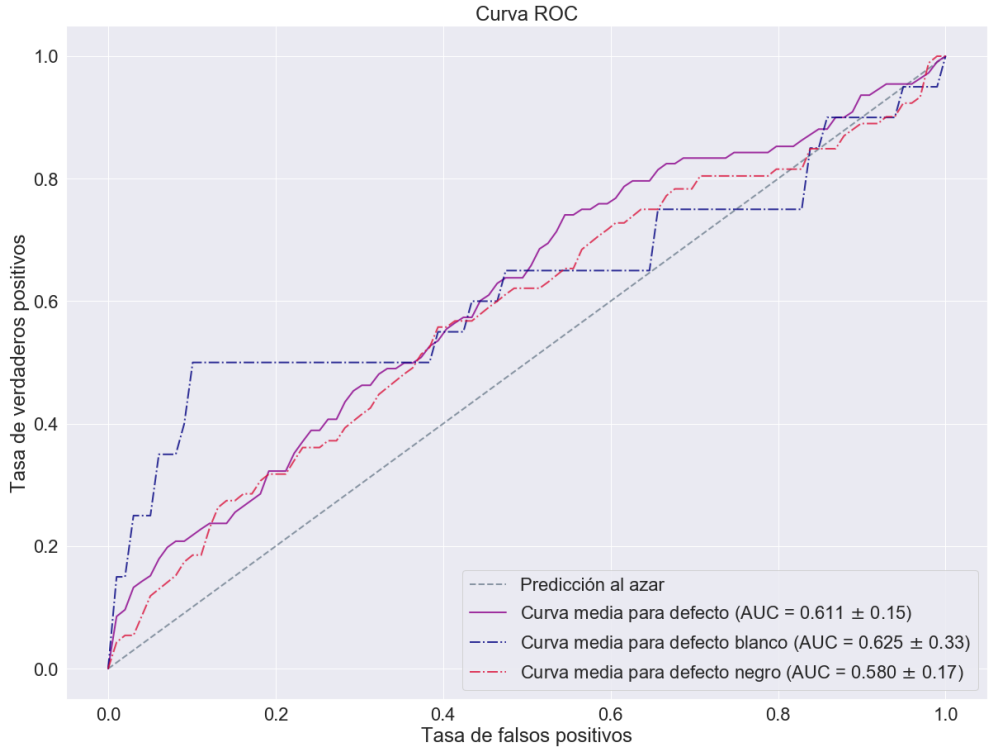

Figura 3.15: Curvas ROC de discriminante lineal empleando el conjunto de datos A del Corpus II.

Resultados de la discriminante lineal en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.422 ± 0.12	0.611 ± 0.15 {	Negro: 0.580 ± 0.17 Blanco: 0.625 ± 0.33

Tabla 3.14: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante lineal creado a partir del conjunto A de datos del Corpus II.

Sistema de discriminante lineal -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3881	102	3983
	Con defecto	12	5	17
		3893	107	4000

- Tasa de acierto: 97.15 %.
  Intervalo de confianza del 95 %: (95.26, 99.03)
  - Sensibilidad: 4.67 %
  - Especificidad: 99.69 %

Tabla 3.15: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante lineal del Corpus II con el conjunto de datos A.

3.3.1.4. Conjunto de datos B

Modelo óptimo: 41 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorB, AreaErrorN.
- Estadísticos: VarianzaOr, MediaTratadaB, SkewnessTratadaB, KurtosisTratadaB, MediaTratadaN, VarianzaTratadaN, SkewnessTratadaN, KurtosisTratadaN.
- DCT: imagen sin binarizar: 6; imagen binarizada: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30.

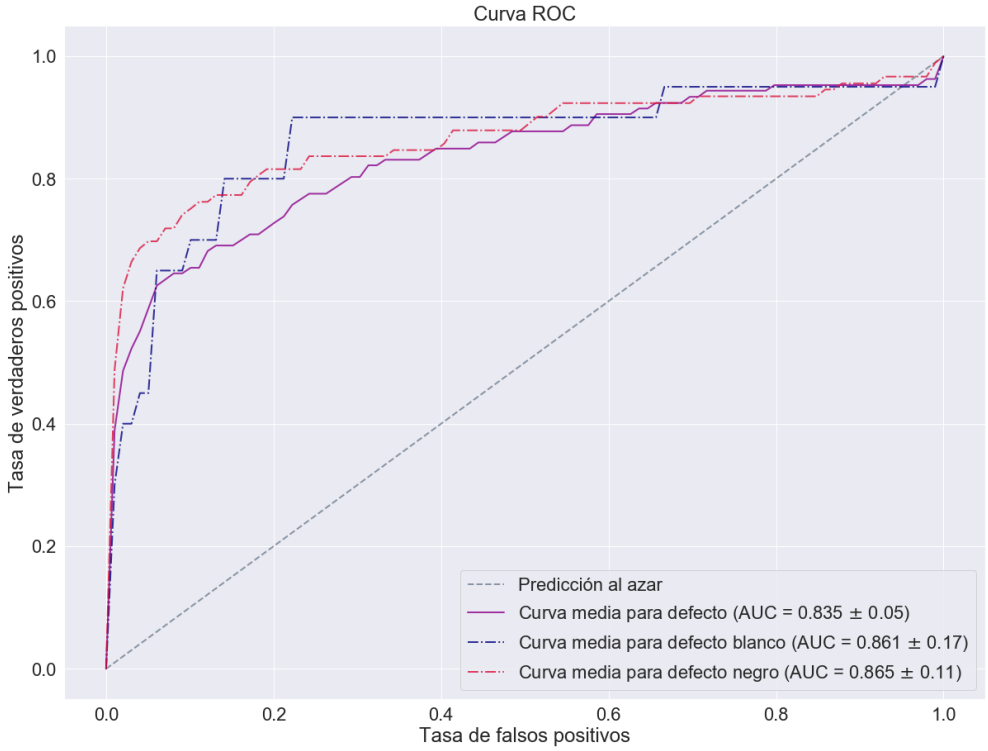

Figura 3.16: Curvas ROC de discriminante lineal empleando el conjunto de datos B del Corpus II.

Resultados de la discriminante lineal en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.254 ± 0.10	0.835 ± 0.05 {	Negro: 0.865 ± 0.11 Blanco: 0.861 ± 0.17

Tabla 3.16: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante lineal creado a partir del conjunto B de datos del Corpus II.

Las curvas ROC (Figura 3.16) así como las estimaciones de tasa de equierror y área bajo la curva (Tabla 3.16) nos indican que este modelo podría ser una mejora significativa con respecto al sistema de referencia en el Corpus II o al creado con el conjunto de datos A del mismo.

Obtenemos un modelo de mejor precisión que el sistema de referencia y más equilibrado que el del conjunto de datos A: aunque la tasa de acierto sea inferior, se etiquetan correctamente más imágenes con defecto sin que la detección de imágenes carentes defectos sea mala (Tabla 3.17). La sensibilidad es aún bastante mejorable.

Sistema de discriminante lineal -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3837	77	3914
	Con defecto	56	30	86
		3893	107	4000

- Tasa de acierto: 96.67 %.
  Intervalo de confianza del 95 %: (94.64, 98.70)
  - Sensibilidad: 28.03 %
  - Especificidad: 98.56 %

Tabla 3.17: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante lineal del Corpus II con el conjunto de datos B.

3.3.2. Discriminante Cuadrático (QDA)

Corpus I

3.3.2.1. Conjunto de datos A

Modelo óptimo: 35 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorN y AreaErrorB.
- PCA: CompPrinOr1, CompPrinOr3, CompPrinOr4, CompPrinTratB1, CompPrinTratB2, CompPrinTratN1, CompPrinTratN2.
- DCT: imagen sin binarizar: 1, 2, 3, 4, 6, 7, 11, 12, 16, 20, 22, 27, 28, 29; imagen binarizada: 2, 3, 4, 6, 7, 11, 16, 20, 22, 28, 29.

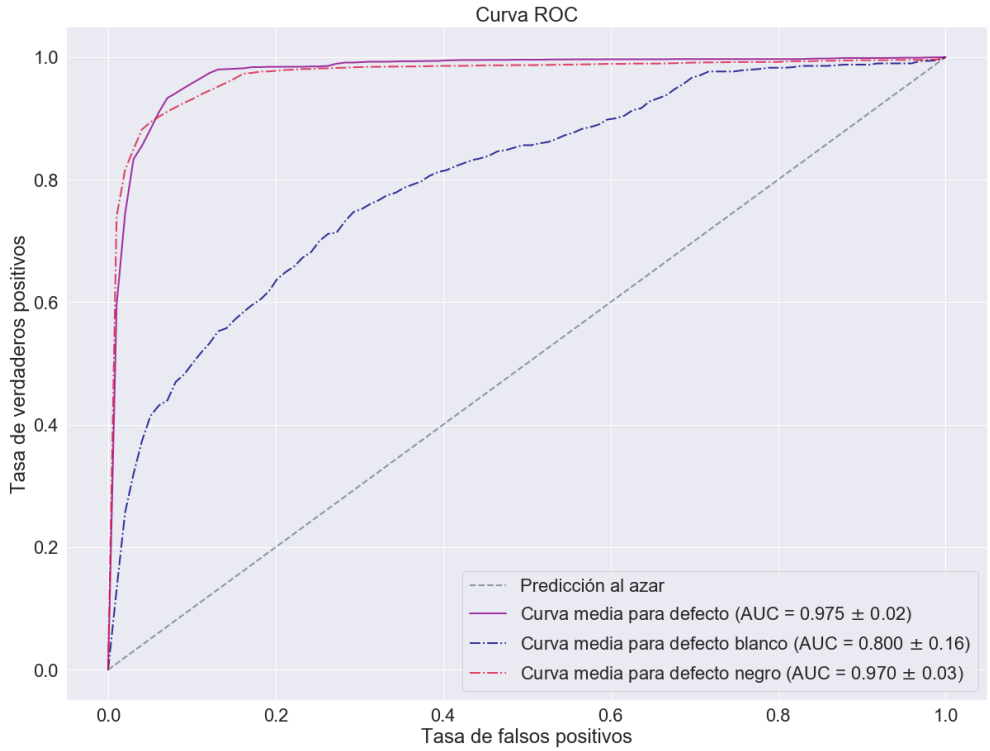

Figura 3.17: Curvas ROC de discriminante cuadrático empleando el conjunto de datos A del Corpus I.

Resultados de la discriminante cuadrático en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.046 ± 0.04	0.975 ± 0.02 {	Negro: 0.970 ± 0.03 Blanco: 0.800 ± 0.16

Tabla 3.18: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante cuadrático creado a partir del conjunto A de datos del Corpus I.

La curva ROC de defecto general así como la de defecto negro se pueden calificar de excelentes (3.17). La correspondiente a los defectos blancos es mejorable. La Tabla 3.18 refleja como el sistema ha obtenido los mejores resultados hasta el momento en el caso del Corpus I.

El modelo con la probabilidad de clasificación asignada da lugar a una tasa de acierto ligeramente mejor que la del sistema de referencia y la obtenida con estos datos en el caso del discriminante lineal (Tabla 3.19). El número de falsos negativos alcanza el valor mínimo obtenido con el sistema de referencia, mientras que el número de falsos positivos es ligeramente superior. Existe un buen equilibrio entre sensibilidad y especificidad.

Sistema de discriminante cuadrático -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	670	252	922
	Con defecto	39	6206	6245
		709	6458	7167

- Tasa de acierto: 95.93 %.
  Intervalo de confianza del 95 %: (94.27, 97.61)
  - Sensibilidad: 96.09 %
  - Especificidad: 94.50 %

Tabla 3.19: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante cuadrático del Corpus I con el conjunto de datos A.

3.3.2.2. Conjunto de datos B

Modelo óptimo: 41 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorB y AreaErrorN.
- Estadísticos: MediaOr, VarianzaOr, KurtosisOr, MedianaOr, P10Or, P90Or, MediaTratadaB, VarianzaTratadaB, SkewnessTratadaB, KurtosisTratadaB, P10TratadaB, P90TratadaB, MediaTratadaN, VarianzaTratadaN, SkewnessTratadaN, KurtosisTratadaN, P10TratadaN P90TratadaN.
- DCT: imagen sin binarizar: 1, 2, 3, 4, 6, 11, 12, 15, 20, 22, 28, 29; imagen binarizada: 1, 2, 3, 4, 11, 15, 20, 22, 28, 29.

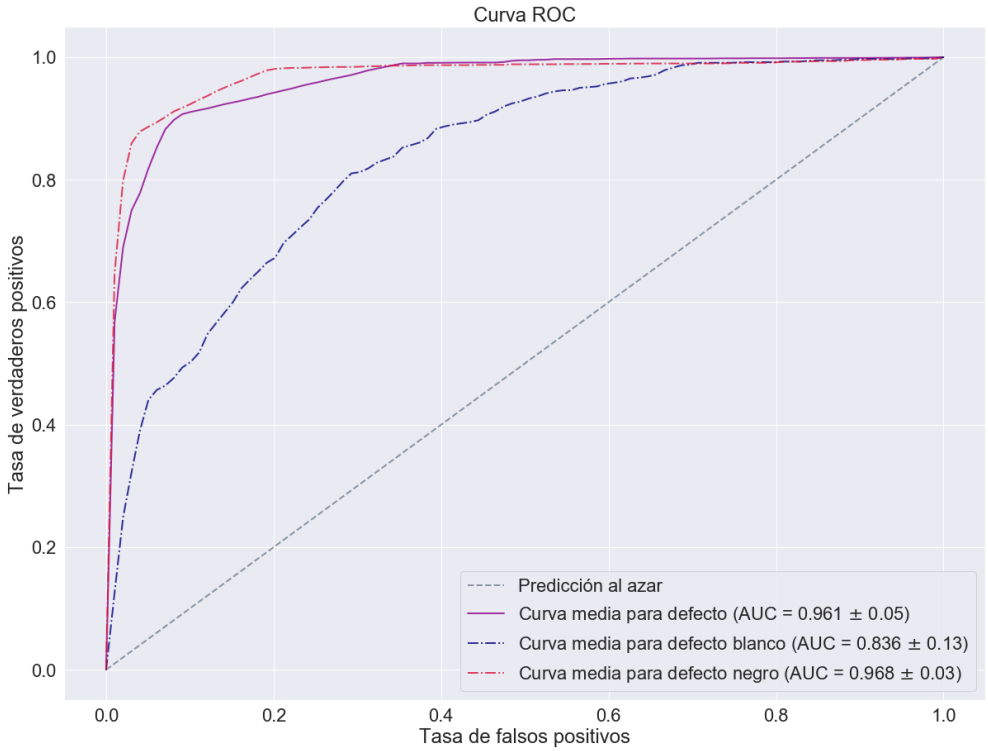

Figura 3.18: Curvas ROC de discriminante cuadrático empleando el conjunto de datos B del Corpus I.

Resultados de la discriminante cuadrático en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.077 ± 0.09	0.961 ± 0.06 {	Negro: 0.968 ± 0.03 Blanco: 0.836 ± 0.13

Tabla 3.20: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante cuadrático creado a partir del conjunto B de datos del Corpus I.

Tanto el valor del área bajo la curva como la tasa de equierror son indicio de que el ajuste cuadrático realizado logra resultados similares a los del sistema de referencia, siendo mejores que los del discriminante lineal (Tabla 3.20). En la Figura 3.18 podemos ver como la detección de defectos blancos es algo peor que la de negros.

Si ajustamos el umbral de clasificación, las conclusiones anteriores son evidenciadas de nuevo. Obtenemos resultados parecidos al conjunto A (Tabla 3.21): se mejora el resultado obtenido con LDA. Se supera también la precisión del sistema de referencia principalmente gracias al agudo descenso de casos de falsos positivos. El número de falsos negativos ha aumentado pero el equilibrio entre especificidad y sensibilidad es adecuado.

Sistema de discriminante cuadrático -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	661	272	933
	Con defecto	48	6186	6234
		709	6458	7167

- Tasa de acierto: 95.53 %.
  Intervalo de confianza del 95 %: (93.79, 97.27)
  - Sensibilidad: 95.79 %
  - Especificidad: 93.23 %

Tabla 3.21: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante cuadrático del Corpus I con el conjunto de datos B.

Corpus II

En el caso de este conjunto de datos se han obtenido clasificadores con serios problemas de exceso de varianza. Por ello, y tal como se explico en el capítulo de la metodología correspondiente, se han hecho dos procedimientos de selección de variables, uno basado en el VIF y otro en una selección de variables recursiva. Esto se debe principalmente a la fuerte correlación entre las variables de las componentes DCT, especialmente en aquellas procedentes de la imagen binarizada.

3.3.2.3. Conjunto de datos A

Comenzamos eliminando las variables cuyo VIF fuese elevado (de 83 variables pasamos a 32). Depués realizamos la eliminación de variables mediante RFE (de 32 variables a 15). **Modelo óptimo**: 15 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorB, AreaErrorN.
- PCA: CompPrinOr1, CompPrinOr3, CompPrinOr4, CompPrinTratB2, CompPrinTratN2.
- DCT: imagen sin binarizar: 2, 3, 4, 5, 7, 14, 15, 16;

Resultados de la discriminante cuadrático en el Corpus II

	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.314 ± 0.14	0.664 ± 0.13 {	Negro: 0.649 ± 0.22 Blanco: 0.500 ± 0.00

Tabla 3.22: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante cuadrático creado a partir del conjunto A de datos del Corpus II.

Puede que el funcionamiento general del sistema haya mejorado si prestamos atención a la tasa de equierror y al área bajo la curva ROC de defecto general (Figura 3.19 y Tabla 3.22), sin embargo la detección de defectos blancos es peor que la del discriminante lineal. Es una asignación completamente azarosa.

Se obtiene un modelo cuya precisión es ligeramente inferior al modelo LDA creado, como podemos observar en la Tabla 3.23. Su sensibilidad es paupérrima (menor que el 2 %) y peor que la del modelo correspondiente de LDA y el sistema de referencia.

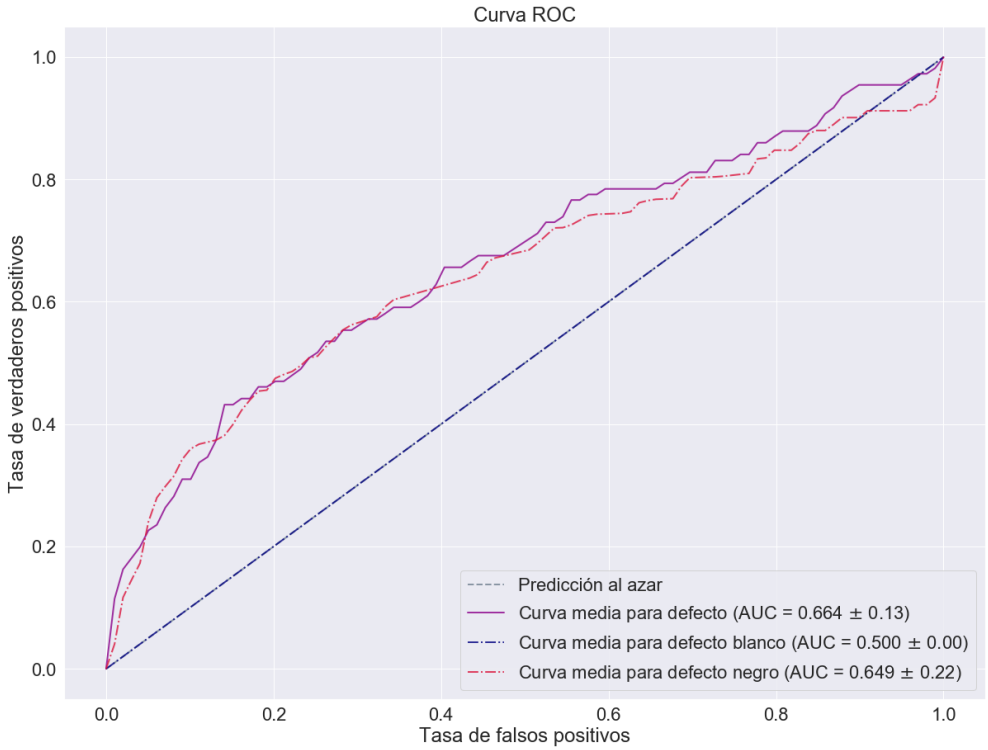

Figura 3.19: Curvas ROC de discriminante cuadrático empleando el conjunto de datos A del Corpus II.

Sistema de discriminante cuadrático -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3872	105	3977
	Con defecto	21	2	23
		3893	107	4000

- Tasa de acierto: 96.85 %.
  Intervalo de confianza del 95 %: (94.87, 98.83)
  - Sensibilidad: 1.87 %
  - Especificidad: 99.46 %

Tabla 3.23: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante cuadrático del Corpus II con el conjunto de datos A.

3.3.2.4. Conjunto de datos B

Primero se han eliminado las variables cuyo VIF fuese elevado (pasando de 70 variables a 28) para posteriormente realizar la eliminación de variables mediante RFE (de 28 variables a 14). **Modelo óptimo**: 14 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- AreaErrorB.
- Estadísticos: SkewnessOr.
- DCT: imagen sin binarizar: 3, 8, 12, 13, 14, 17, 19, 21, 24, 26, 27, 29;

Resultados de la discriminante cuadrático en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.412 ± 0.11	0.589 ± 0.13 {	Negro: 0.600 ± 0.18 Blanco: 0.580 ± 0.16

Tabla 3.24: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de discriminante cuadrático creado a partir del conjunto B de datos del Corpus II.

Las curvas ROC (Figura 3.20) suponen una mejora en lo que se refiere al equilibrio en la detección de defectos blancos y negros con respecto al ajuste con el conjunto A. La Tabla 3.24 caracteriza a un clasificador aparentemente de peor funcionamiento que el creado a partir de estos datos con un ajuste lineal.

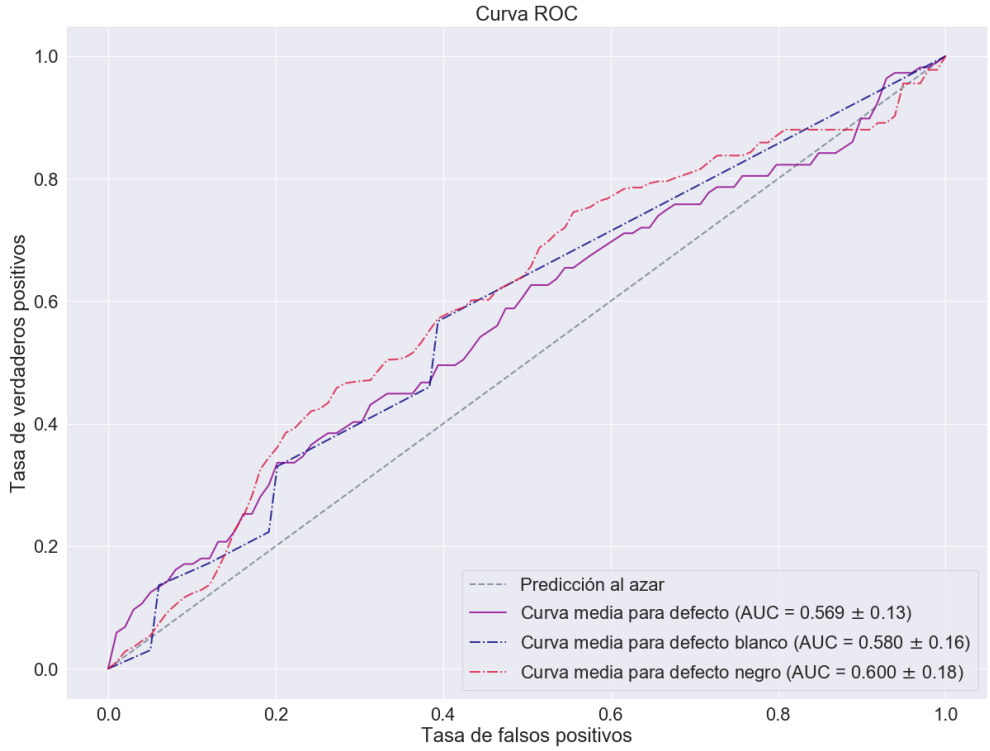

Figura 3.20: Curvas ROC de discriminante cuadrático empleando el conjunto de datos B del Corpus II.

Tal y como podemos ver en la Tabla 3.25, el clasificador obtenido mejora la precisión del sistema de referencia y de los modelos correspondientes de discriminante lineal. Sin embargo, el número de tableros que clasifica como defectuosos es excesivamente bajo, dando lugar a una sensibilidad muy pobre. Se puede destacar como es un clasificador de dimensionalidad bastante baja (14 variables).

Sistema de discriminante cuadrático -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3889	105	3994
	Con defecto	4	2	6
		3893	107	4000

- Tasa de acierto: 97.27 %.
  Intervalo de confianza del 95 %: (95.43, 99.11)
  - Sensibilidad: 1.87 %
  - Especificidad: 99.90 %

Tabla 3.25: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de discriminante cuadrático del Corpus II con el conjunto de datos B.

La Tabla resumen 3.26 nos reincide en que el discriminante lineal no mejora el resultado del sistema de referencia en el caso del Corpus I usando como criterio tanto el valor de la tasa de acierto, tasa de equierror o área bajo la curva. En el Corpus II se mejoran los valores obtenidos a partir del sistema de referencia a pesar de que la sensibilidad de los sistemas diste de ser adecuada.

La mayor flexibilidad del discriminante cuadrático supone una mejora leve con respecto al sistema de referencia y al LDA en la mayoría de casos si atendemos a las mediciones estimadas de la tasa de acierto, área bajo la curva y EER. Cabe a destacar la baja dimensionalidad de los modelos del Corpus II. Estos carecen de utilidad práctica ya que carecen de una sensibilidad adecuada.

Tasa de acierto estimada de los modelos análisis discriminante								
		Corpus I				Corpus II		
		Variables	EER	AUC	Tasa de acierto	Variables	EER	AUC	Tasa de acierto
LDA	Conjunto A	35	0.103	0.934	95.29 %	35	0.422	0.611	97.15 %
	Conjunto B	41	0.100	0.944	94.08 %	41	0.254	0.835	96.67 %
QDA	Conjunto A	35	0.046	0.975	95.94 %	15	0.314	0.664	96.85 %
	Conjunto B	41	0.077	0.961	95.53 %	14	0.412	0.589	97.27 %

Tabla 3.26: Número de variables, AUC y EER, así como tasa de acierto estimada por validación cruzada del modelo seleccionado, de los sistemas de discriminante lineal (LDA) y cuadrático (QDA) haciendo uso de los conjuntos de datos A y B en ambos Corpus.

3.4. Árboles de decisión

A lo largo de la siguiente sección se han desarrollado modelos de la siguiente naturaleza:

- Árboles de decisión básicos estilo C4.5.
- Métodos de asociación de árboles: Bagging (Random Forest) y Boosting (AdaBoost y BART).

3.4.1. Árboles de decisión estilo C4.5

De acuerdo a la metodología establecida en la sección 2.3 se han desarrollado los árboles de decisión estilo C4.5. Se han ajustado por validación cruzada, tal y como se comento en la metodología, los parámetros C y M que controlan el podado y el número de instancias usadas en cada nodo respectivamente.

Corpus I

3.4.1.1. Conjunto de datos A

Modelo óptimo: 20 variables. Seleccionado bajo criterio de máxima tasa de acierto estimada por validación cruzada. Parámetros: $M = 1, C = 0.01$.

- PCA: CompPrinOr1, CompPrinOr4, CompPrinTratB1, CompPrinTratB2, CompPrinTratN2.
- DCT: imagen sin binarizar: 1, 2, 3, 6, 21, 23; imagen binarizada: 1, 6, 7, 10, 15, 16, 18, 27, 28.

La Figura 3.21 es la visualización del árbol de decisión inducido. Éste emplea en nodos de decisión las variables DCT en numerosas ocasiones. Los atributos relacionados con el área de error no son usados.

Resultados del árbol de decisión estilo C4.5 en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.090 ± 0.03	0.940 ± 0.01 {	Negro: 0.804 ± 0.01 Blanco: 0.521 ± 0.05

Tabla 3.27: Estimación de la tasa de equierror y el área bajo la curva ROC del árbol de decisión estilo C4.5 creado a partir del conjunto A de datos del Corpus I.

Las curvas ROC son bastante dispares (Figura 3.22): la de defecto general es bastante buena, mientras que la correspondiente al defecto negro es aceptable y la del blanco es similar a la decisión al azar. La tasa de equierror (Tabla 3.27) es algo mejor que la de los modelos LDA creados anteriormente.

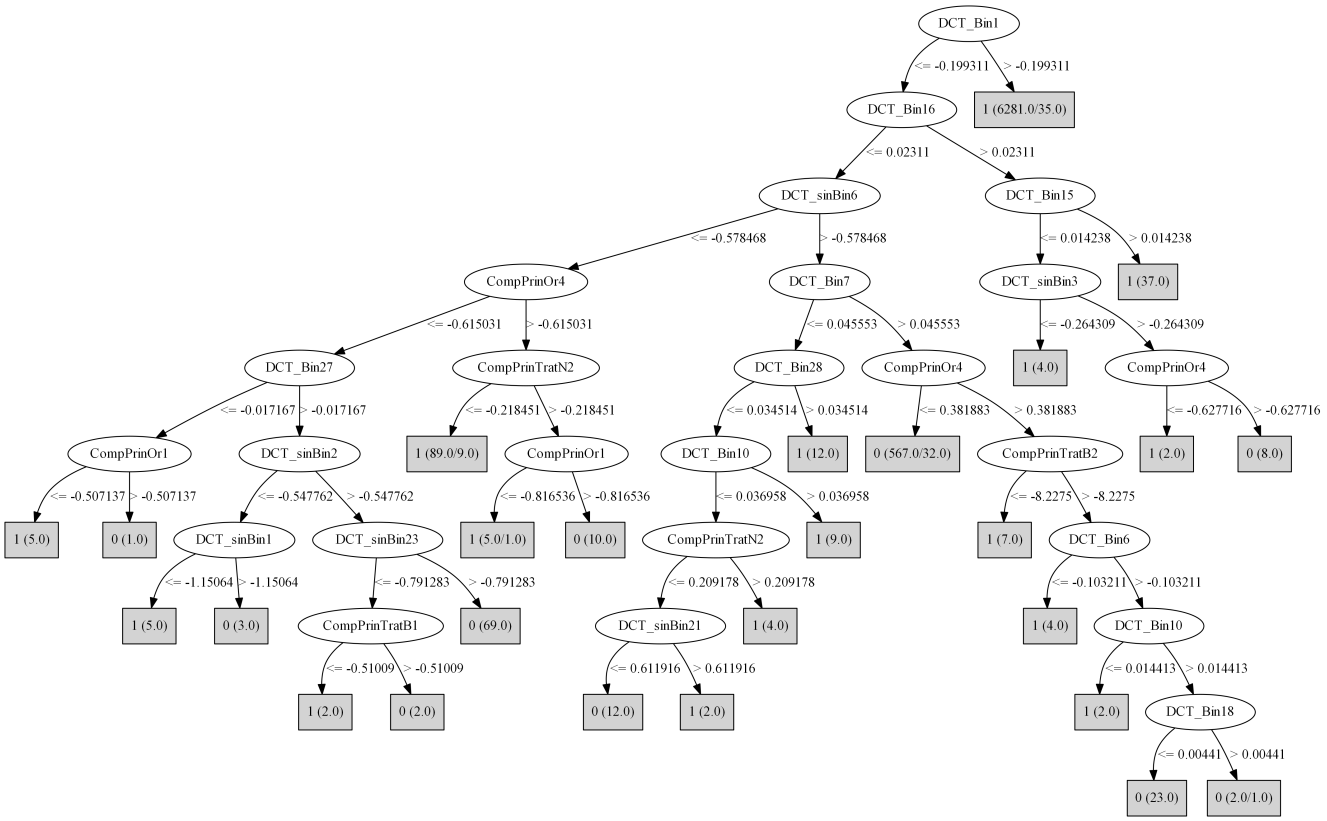

Figura 3.21: Árbol de decisión estilo C4.5 inducido a partir del conjunto de datos A del Corpus I.

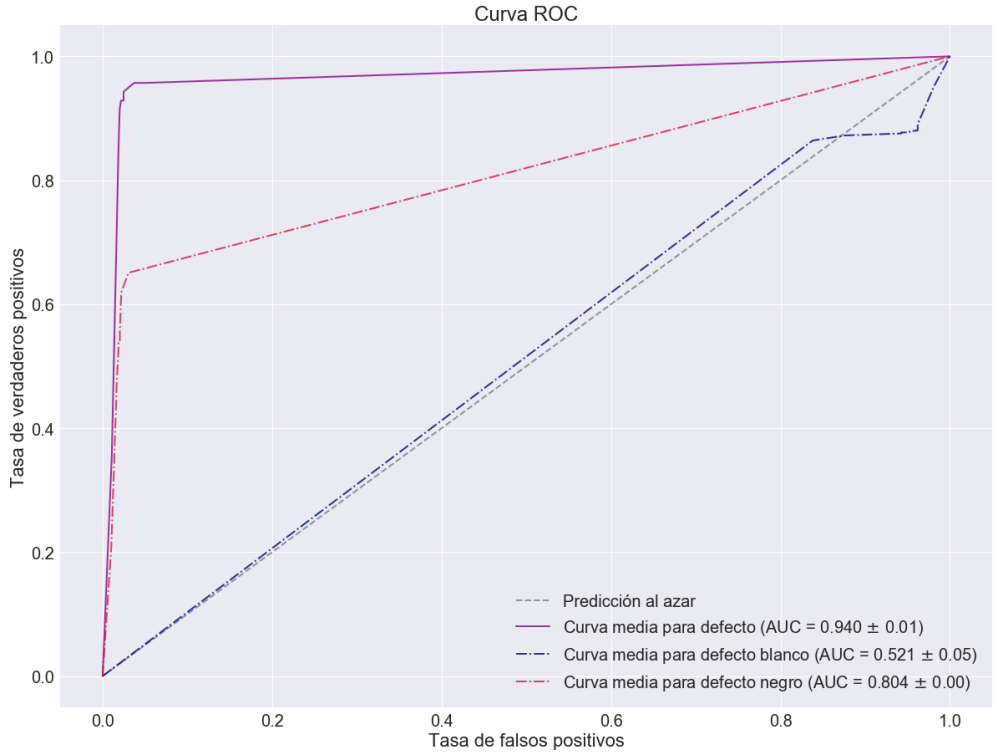

Figura 3.22: Curvas ROC del árbol de decisión estilo C4.5 empleando el conjunto de datos A del Corpus I.

La tasa de acierto del clasificador es la más elevada hasta el momento en el caso del Corpus I (Tabla 3.28). Tiene características propias de ser un buen clasificador: tiene una elevada tasa de acierto, es de baja dimensionalidad y elevados valores de la especificidad y sensibilidad. Hay cierta disparidad entre las dos.

Árbol de decisión estilo C4.5 -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	630	72	702
	Con defecto	79	6386	6465
		709	6458	7167

- Tasa de acierto: 97.89 %.
  Intervalo de confianza del 95 %: (96.67, 99.10)
  - Sensibilidad: 98.88 %
  - Especificidad: 88.85 %

Tabla 3.28: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de árbol de decisión estilo C4.5 del Corpus I con el conjunto de datos A.

3.4.1.2. Conjunto de datos B

Modelo óptimo: 16 variables. Seleccionado bajo criterio de máxima tasa de acierto estimada por validación cruzada. Parámetros: $M = 3, C = 0.01$.

- AreaErrorN.
- Estadísticos: MediaOr, KurtosisOr, MedianaOr, MediaTratadaN, VarianzaTratadaN, SkewnessTratadaN, KurtosisTratadaN, P10TratadaN.
- DCT: imagen sin binarizar: 6, 29; imagen binarizada: 1, 6, 10, 19, 29.

La Figura 3.24 es la visualización del árbol de decisión inducido. Las variables relacionadas con el filtrado de defectos negros y las componentes de la DCT extraídas de la imagen binarizada son relevantes.

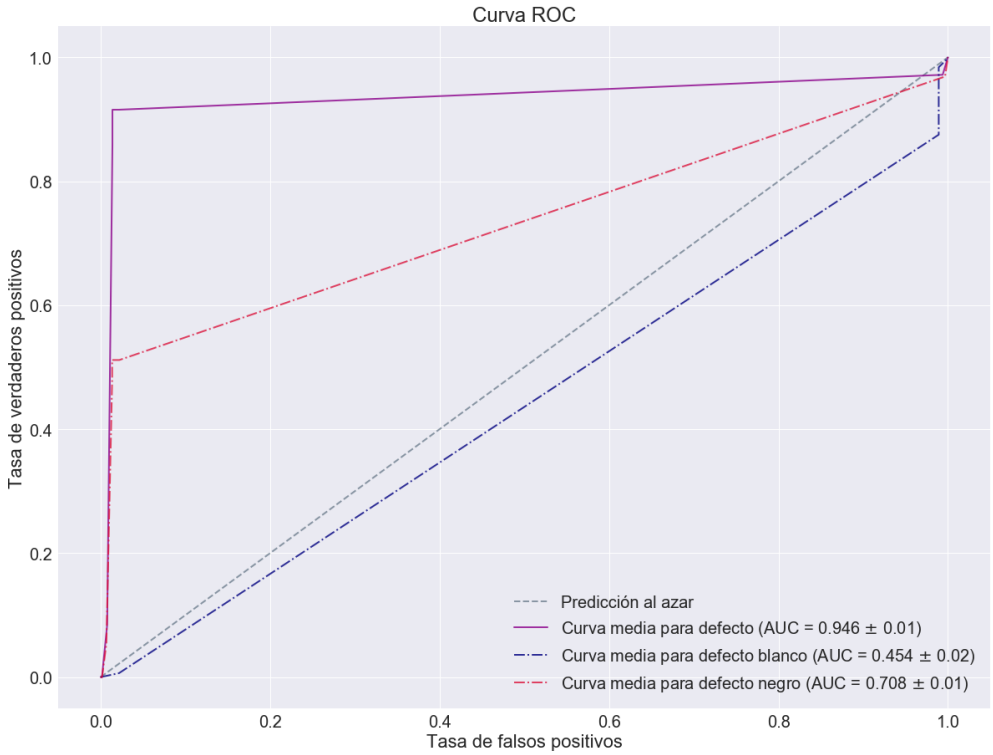

Figura 3.23: Curvas ROC del árbol de decisión estilo C4.5 empleando el conjunto de datos B del Corpus I.

Las curvas ROC son aún más dispares que en el caso del conjunto de datos A (Figura 3.23). El que el árbol no haya creado ningún nodo de decisión a partir de las variables propias del filtrado de blancos podría estar relacionado con que la discriminación de blancos sea peor que la asignación azarosa. La tasa de equierror (Tabla 3.29) es inferior que la del árbol del conjunto de datos A y que los modelos LDA anteriores.

Resultados del árbol de decisión estilo C4.5 en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.076 ± 0.02	0.946 ± 0.01	Negro: 0.708 ± 0.02 Blanco: 0.454 ± 0.01

Tabla 3.29: Estimación de la tasa de equierror y el área bajo la curva ROC del árbol de decisión estilo C4.5 creado a partir del conjunto B de datos del Corpus I.

Los resultados al ajustar el umbral probabilístico de clasificación son similares a los del conjunto A (Tabla 3.30). Supera su precisión (y por tanto la de los sistemas de referencia y los modelos de análisis discriminante) con menos variables y mejora ligeramente la especificidad.

Árbol de decisión estilo C4.5 -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	637	65	703
	Con defecto	72	6393	6465
		709	6458	7167

- Tasa de acierto: 98.09 %.
  Intervalo de confianza del 95 %: (96.93, 99.25)
  - Sensibilidad: 98.99 %
  - Especificidad: 89.84 %

Tabla 3.30: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de árbol de decisión estilo C4.5 del Corpus I con el conjunto de datos B.

Corpus II

Para los modelo del Corpus II ha sido necesario hacer un muestreo tal que el la mitad de las observaciones de imágenes defectuosas se hayan añadido de nuevo al conjunto de datos para dar más peso a estas en la inducción del clasificador.

3.4.1.3. Conjunto de datos A

Modelo óptimo: 31 variables. Seleccionado bajo criterio de máxima área bajo la curva ROC estimada por validación cruzada. Parámetros: $M = 2, C = 0.31$.

- PCA: CompPrinOr1, CompPrinOr2, CompPrinOr3, CompPrinOr4, CompPrinTratB1, CompPrinTratB2, CompPrinTratN1.
- DCT: imagen sin binarizar: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 17, 21, 23, 24, 26, 27, 28, 30; imagen binarizada: 6, 8.

La Figura 3.26 es la visualización del árbol de decisión inducido. Se trata de un árbol muy denso en el que el número de variables requerido es más elevado que en el Corpus I. Las variables relacionadas con las componentes de la DCT de la imagen sin binarizar parecen ser relevantes, necesitándose 22 sobre el total de 30 de éstas.

Resultados del árbol de decisión estilo C4.5 en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.488 ± 0.35	0.596 ± 0.02	Negro: 0.588 ± 0.02 Blanco: 0.677 ± 0.04

Tabla 3.31: Estimación de la tasa de equierror y el área bajo la curva ROC del árbol de decisión estilo C4.5 creado a partir del conjunto A de datos del Corpus II.

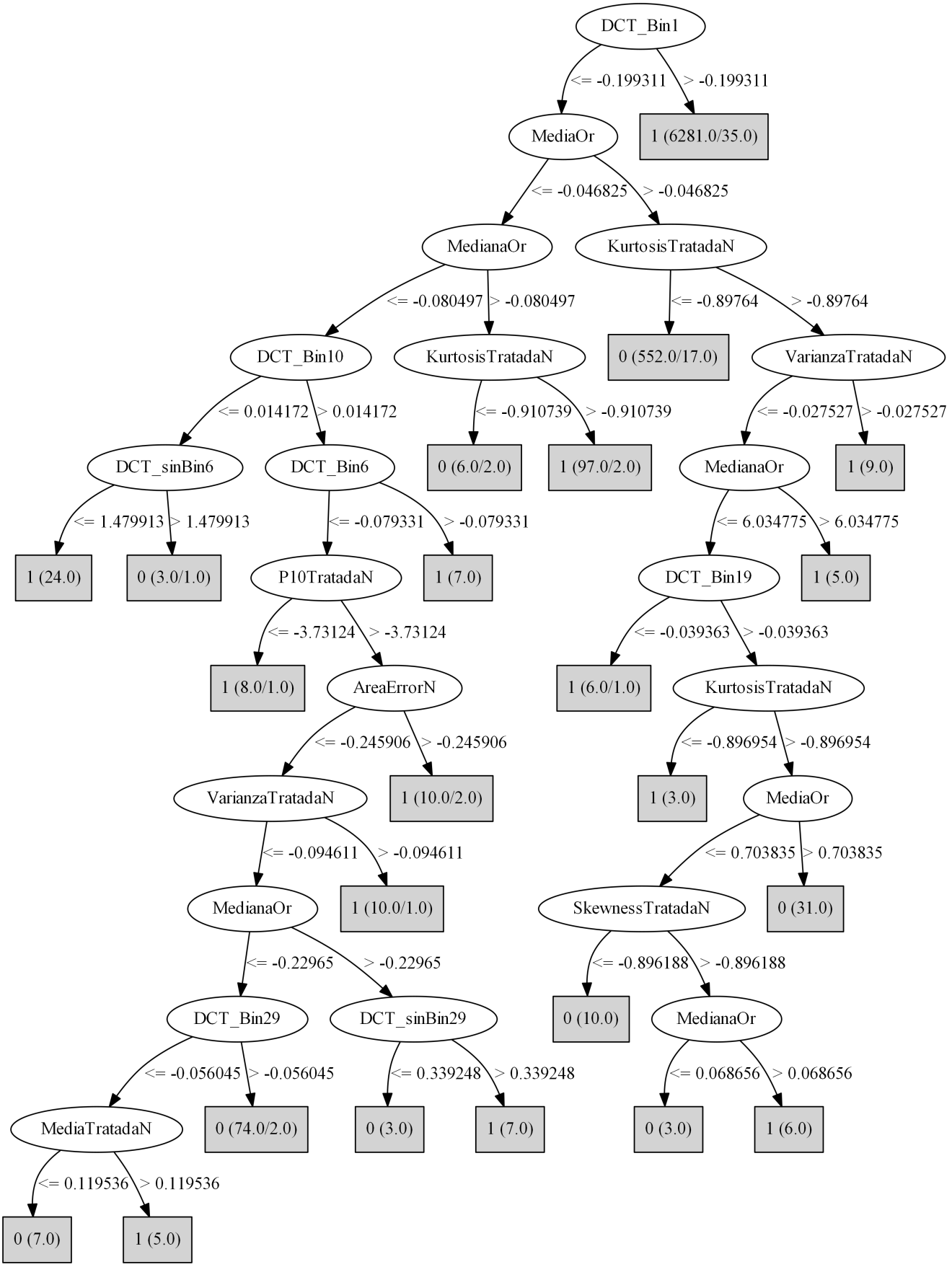

Figura 3.24: Árbol de decisión estilo C4.5 inducido a partir del conjunto de datos B del Corpus I.

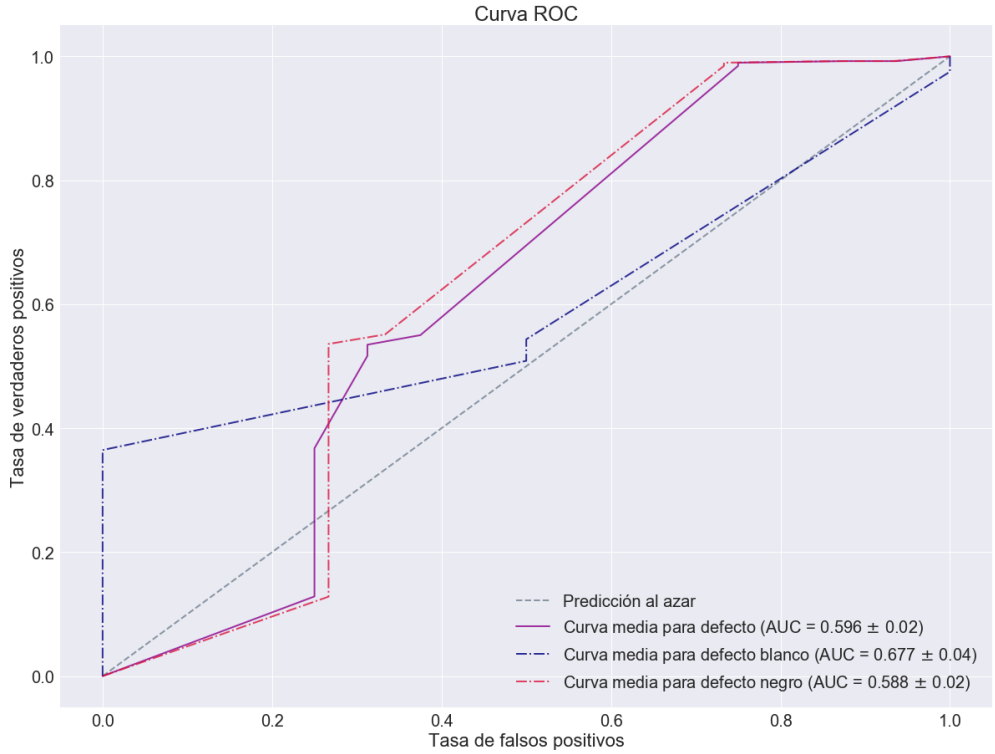

Figura 3.25: Curvas ROC del árbol de decisión estilo C4.5 empleando el conjunto de datos A del Corpus II.

La Figura 3.25 muestra las curvas ROC asociadas al árbol de decisión inducido. Son irregulares y su área por debajo no es muy elevada. La tasa de equierror (Tabla 3.27) es muy variable y elevada, lo que nos indica que el clasificador posiblemente no tenga una buena capacidad discriminatoria de clases.

El modelo con el umbral de clasificación ajustado da lugar a una tasa de acierto bastante adecuada (Tabla 3.32). La sensibilidad tiene un valor superior al de los modelos de análisis discriminante por lo que podría ser indicativo de que el clasificador es más equilibrado.

Árbol de decisión estilo C4.5 -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3814	114	3928
	Con defecto	80	46	126
		3893	160	4053

- Tasa de acierto: 95.24 %.
  Intervalo de confianza del 95 %: (92.83, 97.65)
  - Sensibilidad: 28.75 %
  - Especificidad: 97.97 %

Tabla 3.32: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de árbol de decisión estilo C4.5 del Corpus II con el conjunto de datos A.

3.4.1.4. Conjunto de datos B

Modelo óptimo: 38 variables. Seleccionado bajo criterio de máxima área bajo la curva ROC estimada por validación cruzada. Parámetros: $M = 1, C = 0.46$.

- AreaErrorN.
- Estadísticos: MediaOr, VarianzaOr, SkewnessOr, KurtosisOr, MedianaOr, P10Or, P90Or, MediaTratadaB, SkewnessTratadaB, KurtosisTratadaB, MedianaTratadaB, P10TratadaB, P90TratadaB, MediaTratadaN, VarianzaTratadaN, SkewnessTratadaN, KurtosisTratadaN.
- DCT: imagen sin binarizar: 1, 2, 3, 6, 8, 10, 11, 12, 14, 16, 18, 19, 21, 23, 24, 27, 29, 30; imagen binarizada: 3, 4.

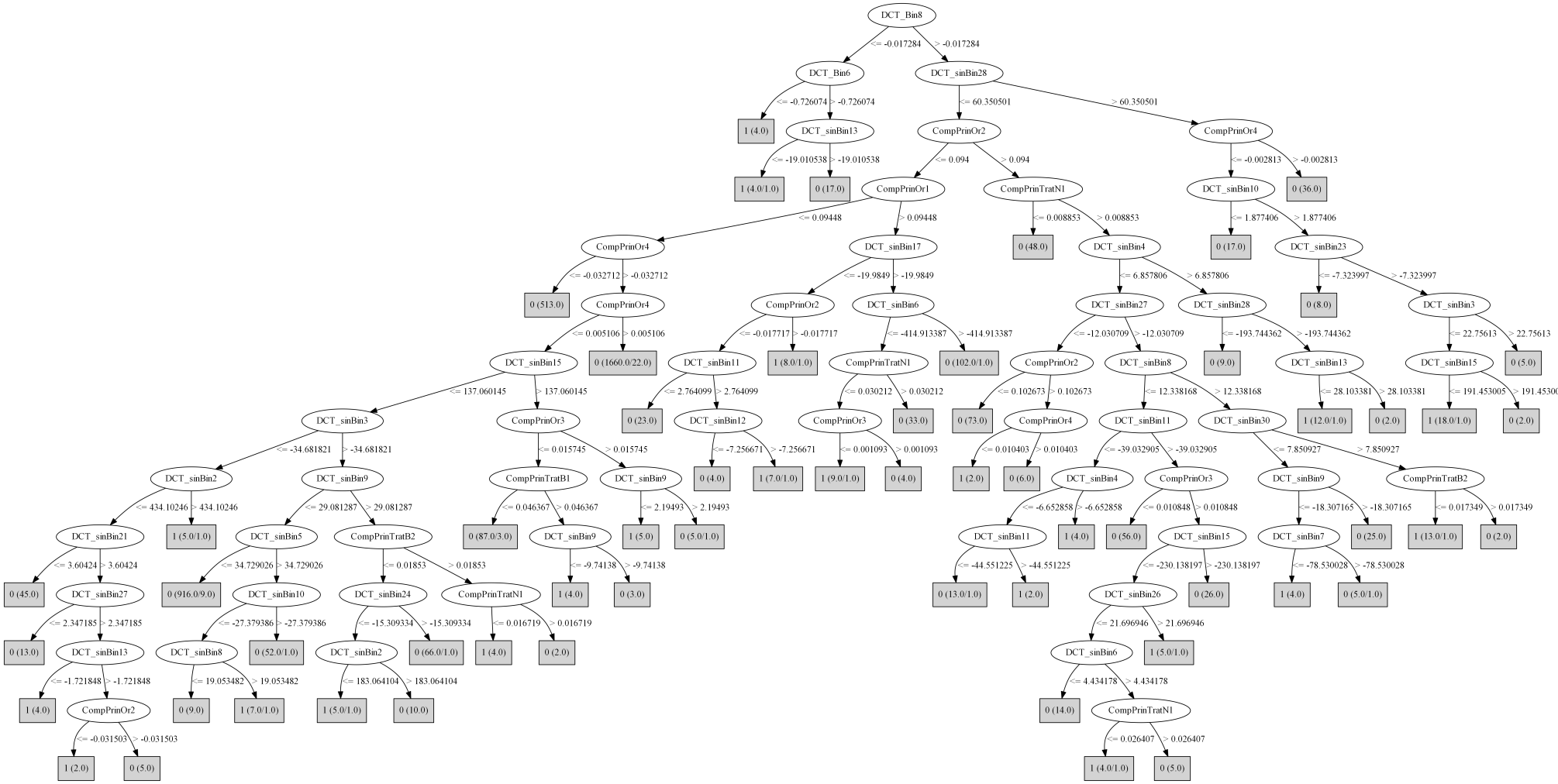

Figura 3.26: Árbol de decisión estilo C4.5 inducido a partir del conjunto de datos A del Corpus II.

La Figura 3.28 es la visualización del árbol de decisión inducido. Se le otorga mucha importancia a las componentes de la DCT extraídas de la imagen binarizada así como a la variable KurtosisTratadaN. Esta variable aparece en varios nodos de decisión, entre ellos, en la raíz.

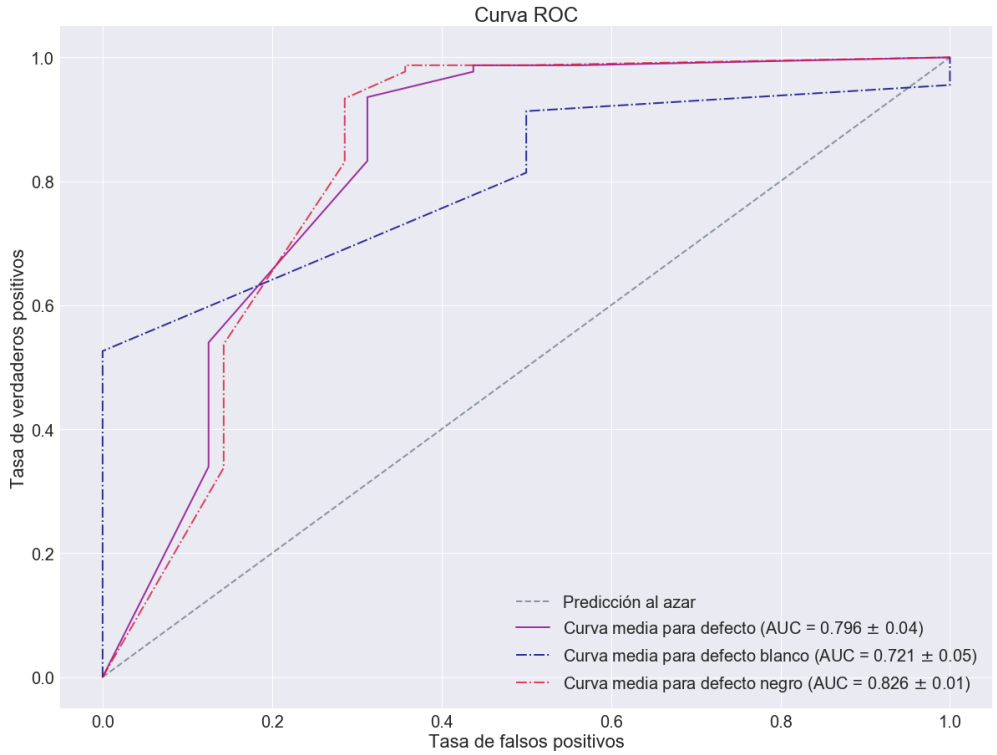

Figura 3.27: Curvas ROC del árbol de decisión estilo C4.5 empleando el conjunto de datos B del Corpus II.

Resultados del árbol de decisión estilo C4.5 en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.196 ± 0.10	0.796 ± 0.04	Negro: 0.826 ± 0.01 Blanco: 0.721 ± 0.05

Tabla 3.33: Estimación de la tasa de equierror y el área bajo la curva ROC del árbol de decisión estilo C4.5 creado a partir del conjunto B de datos del Corpus II.

Las curvas ROC (Figura 3.27), las medidas relacionadas con el área bajo ellas y la tasa de equierror (Tabla 3.33) nos indican que el modelo resultante es altamente satisfactorio en lo que se refiere al Corpus II. Obtiene la EER más baja de los modelos creados y los valores de las tres áreas bajo la curva ROC están bastante equilibrados.

El modelo con umbral de clasificación asignado reincide en lo altamente satisfactorio que es este modelo con respecto a los anteriores del Corpus II (Tabla 3.34). Obtenemos el primer modelo cuya sensibilidad es superior al 50 %. La tasa de acierto supera el 96.5 %.

Árbol de decisión estilo C4.5 -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3825	66	3891
	Con defecto	68	94	162
		3893	160	4053

- Tasa de acierto: 96.69 %.
  Intervalo de confianza del 95 %: (94.66, 98.71)
  - Sensibilidad: 58.75 %
  - Especificidad: 98.25 %

Tabla 3.34: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de árbol de decisión estilo C4.5 del Corpus II con el conjunto de datos B.

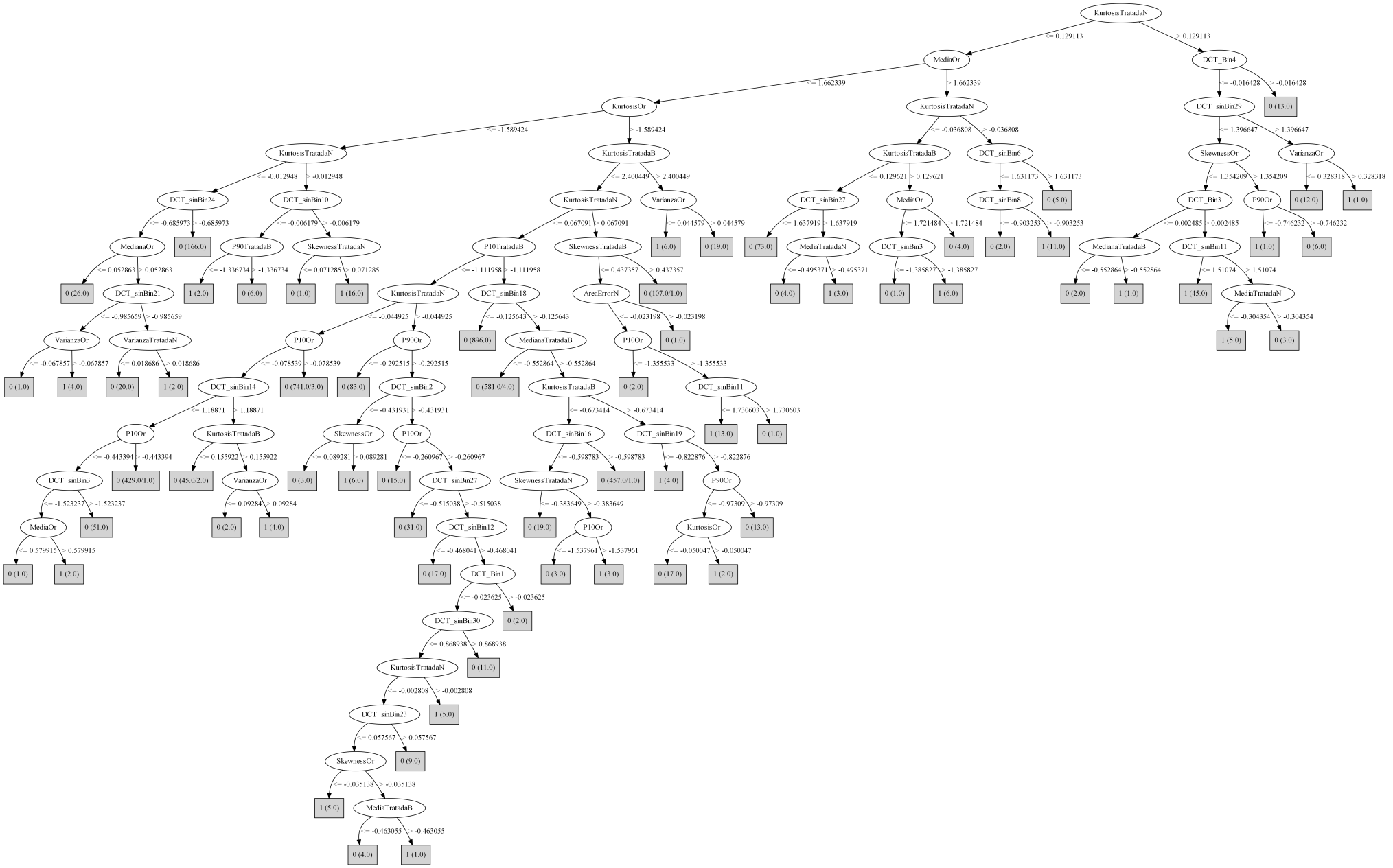

Figura 3.28: Árbol de decisión estilo C4.5 inducido a partir del conjunto de datos B del Corpus II.

La Tabla resumen 3.35 muestra los resultados asociados a los árboles de decisión inducidos. Son altamente satisfactorios tanto por su precisión como por la baja dimensionalidad de estos.

En el caso del Corpus I se obtienen medidas (área bajo la curva ROC, tasa de equierror, tasa de acierto) que apuntan a la superioridad de los modelos arbóreos C4.5 con respecto al sistema de referencia y al discriminante cuadrático y lineal. Esto se logra con un subconjunto de variables muy reducido, no llegándose a emplear ni el 30 % de las variables de entrada al clasificador. El conjunto B en este Corpus parece dar lugar a un árbol de decisión que obtiene mejores resultados.

En el Corpus II han sido necesarias un mayor número de variables. Esto se debe posiblemente por la abundancia de ruido en las imágenes del Corpus II al ser de menor resolución que las del Corpus I. El árbol de decisión creado a partir del conjunto B de variables supera en todos los sentidos los resultados del A así como al sistema de referencia y a la mayoría de clasificadores creados previamente.

Resultados de los árboles de decisión estilo C4.5								
	Corpus I				Corpus II			
	Variables	EER	AUC	Tasa de acierto	Variables	EER	AUC	Tasa de acierto
Conjunto A	20	0.090	0.940	97.89 %	31	0.488	0.596	95.24 %
Conjunto B	16	0.076	0.946	98.09 %	38	0.196	0.796	96.69 %

Tabla 3.35: Número de variables, AUC y EER, así como tasa de acierto estimada por validación cruzada del modelo seleccionado, de los árboles de decisión inducidos a partir de los conjuntos de datos A y B en ambos Corpus.

3.4.2. Random Forest

Los modelos de Random Forest se han desarrollado tal y como aparece en la Sección 2.3.1. Se ha realizado una selección de variables en función de cuales aportaban más variabilidad explicada al modelo. También se ha ajustado el parámetro que define el número de variables seleccionadas aleatoriamente para emplear en cada partición.

Corpus I

3.4.2.1. Conjunto de datos A

Modelo óptimo: 16 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada. Número de variables seleccionadas aleatoriamente por cada partición: 1.

- AreaErrorN.
- PCA: CompPrinOr1, CompPrinOr2, CompPrinOr3, CompPrinOr4.
- DCT: imagen sin binarizar: 1, 6, 7, 16, 28, 29; imagen binarizada: 1, 2, 6, 7, 15.

Resultados de Random Forest en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.066 ± 0.02	0.957 ± 0.01 {	Negro: 0.892 ± 0.02 Blanco: 0.609 ± 0.06

Tabla 3.36: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de Random Forest creado a partir del conjunto A de datos del Corpus I.

Las curvas ROC (Figura 3.29) nos sugieren que el clasificador tiene un comportamiento notablemente bueno con los defectos negros y en general aunque la detección de blancos parece estar muy por detrás. La tasa de equierror toma un valor muy bajo reincidiendo en la posible calidad del clasificador (Tabla 3.36).

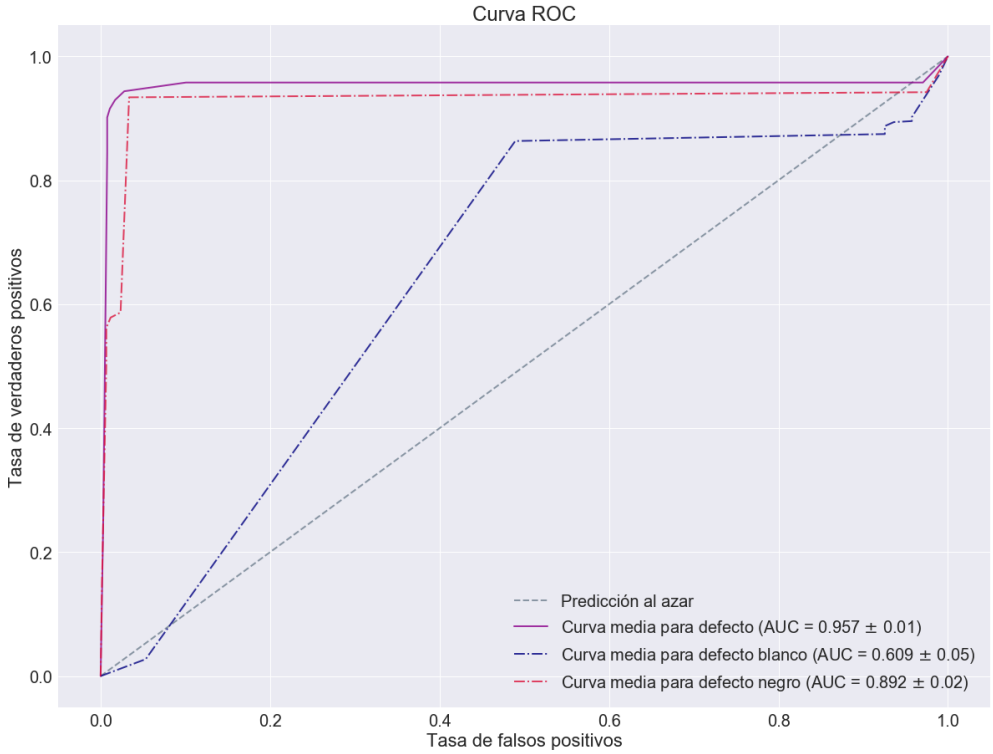

Figura 3.29: Curvas ROC de Random Forest empleando el conjunto de datos A del Corpus I.

El clasificador obtenido es excelente: con tan sólo 16 variables logra una tasa de acierto muy alta (Tabla 3.37). La detección de imágenes con defectos es muy precisa. La detección de tableros sin defectos también es bastante buena aunque admitiría alguna ligera mejora.

Sistema Random Forest -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	651	53	709
	Con defecto	58	6405	6465
		709	6458	7167

- Tasa de acierto: 98.45 %.
  Intervalo de confianza del 95 %: (97.41, 99.49)
  - Sensibilidad: 99.18 %
  - Especificidad: 91.82 %

Tabla 3.37: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de Random Forest del Corpus I con el conjunto de datos A.

3.4.2.2. Conjunto de datos B

Modelo óptimo: 36 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada. Número de variables seleccionadas aleatoriamente por cada partición: 4.

- AreaErrorB y AreaErrorN.
- Estadísticos: MediaOr, P10Or, MedianaOr, SkewnessOr ,P90Or, VarianzaOr, KurtosisOr, KurtosisTratadaB, SkewnessTratadaB, MediaTratadaB, VarianzaTratadaB, KurtosisTratadaN, VarianzaTratadaN, SkewnessTratadaN, MediaTratadaN.
- DCT: imagen sin binarizar: 1, 2, 3, 6, 7, 15, 16, 23, 28, 29; imagen binarizada: 1, 2, 3, 4, 6, 7, 15, 16, 28.

El modelo de Random Forest inducido con el conjunto de variables B logra los mejores valores en cuanto a tasa de equierror y área bajo la curva del defecto general de todos los clasificadores estudiados del Corpus I (Tabla 3.38). En las curvas ROC de la Figura 3.30 podemos ver cómo la detección de defectos en general es excelente aunque la detección de defectos blancos no sea tan extraordinaria.

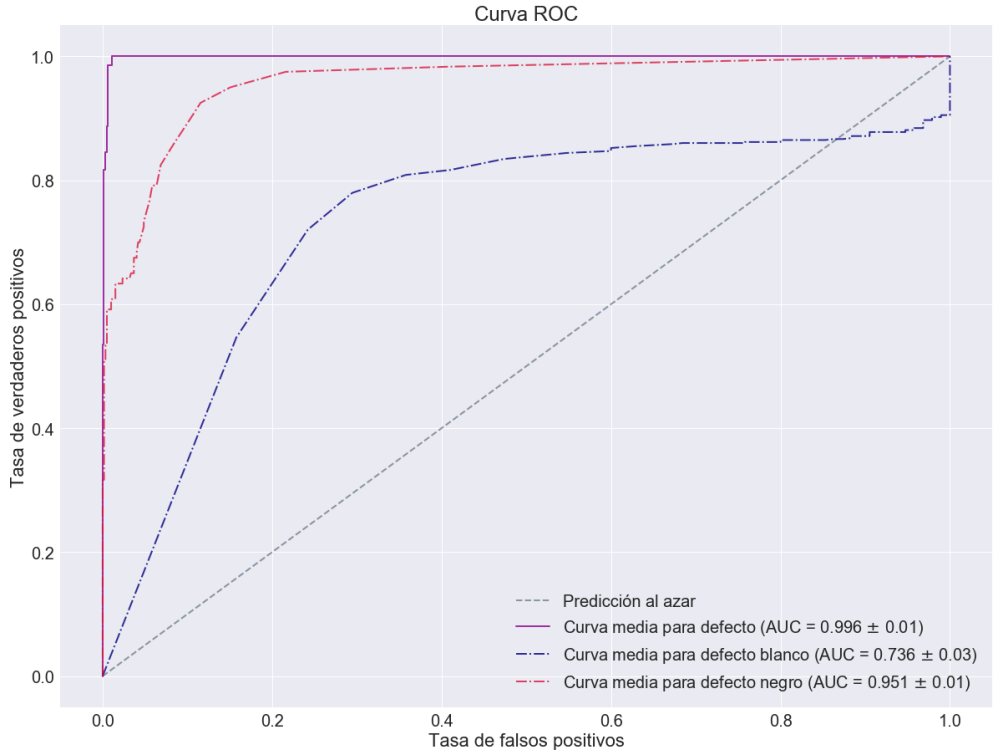

Figura 3.30: Curvas ROC de Random Forest empleando el conjunto de datos B del Corpus I.

Resultados de Random Forest en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.021 ± 0.01	0.996 ± 0.01	Negro: 0.951 ± 0.01 Blanco: 0.736 ± 0.03

Tabla 3.38: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de Random Forest creado a partir del conjunto B de datos del Corpus I.

El clasificador óptimo resultado de la selección de variables con el conjunto B es de una dimensionalidad mayor que la obtenida con el conjunto A, pero logra una aún mejor tasa de acierto que roza el 99 % de precisión (Tabla 3.39). Tanto sensibilidad como especificidad han aumentado con respecto al modelo del conjunto de datos A.

Sistema Random Forest -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	666	43	709
	Con defecto	43	6415	6458
		709	6458	7167

- Tasa de acierto: 98.80 %.
  Intervalo de confianza del 95 %: (97.88, 99.72)
  - Sensibilidad: 99.33 %
  - Especificidad: 93.94 %

Tabla 3.39: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de Random Forest del Corpus I con el conjunto de datos B.

Corpus II

3.4.2.3. Conjunto de datos A

Para este modelo ha sido necesario muestrear las observaciones de imágenes defectuosas de tal forma que la mitad de ellas se añadían al conjunto de datos una vez a mayores. De no hacerse de esta forma, el entrenamiento no se realizaba correctamente y daba lugar a un clasificador muy desequilibrado.

Modelo óptimo: 14 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada. Número de variables seleccionadas aleatoriamente por cada partición: 2.

- PCA: CompPrinOr1, CompPrinOr2, CompPrinOr3, CompPrinOr4, CompPrinTratB2.
- DCT: imagen sin binarizar: 1, 2, 4, 6, 14, 15, 16, 23, 28, 29.

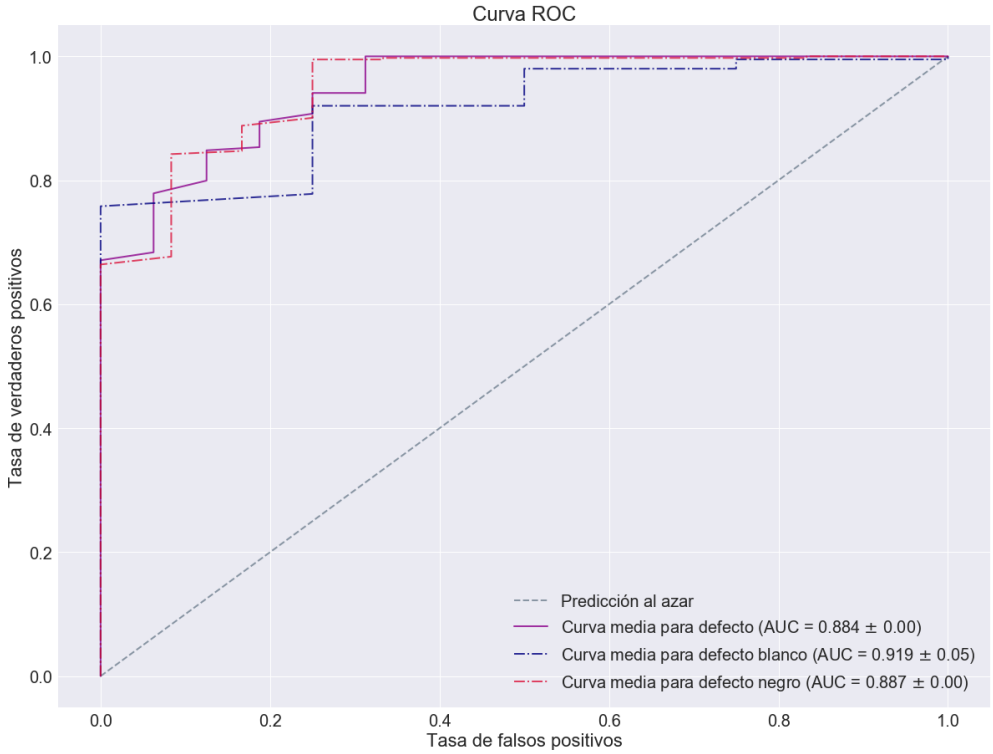

Figura 3.31: Curvas ROC de Random Forest empleando el conjunto de datos A del Corpus II.

Resultados de Random Forest en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.222 ± 0.06	0.884 ± 0.00 {	Negro: 0.887 ± 0.00 Blanco: 0.919 ± 0.05

Tabla 3.40: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de Random Forest creado a partir del conjunto A de datos del Corpus II.

Las estimaciones del área bajo las curvas ROC (Tabla 3.40 y Figura 3.31) nos indican que hemos obtenido un clasificador que puede dar lugar a muy buenos resultados. Además la detección de defectos parece ser equilibrada entre los dos tipos. La tasa de equierror es algo superior al 20 %.

El modelo obtenido tras ajustar la probabilidad de asignación logra una precisión muy elevada y una aparente perfecta capacidad para distinguir tablero sin defectos (Tabla 3.41). Este clasificador tiene un sesgo adicional debido a los pesos aportados a las observaciones, por lo que los resultados pueden tener cierta distorsión. Aún así es un modelo cuya sensibilidad llega al 60 %, más que ninguno de los estudiados en el caso del Corpus II.

Sistema Random Forest -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3893	64	3961
	Con defecto	0	96	92
		3893	160	4053

- Tasa de acierto: 98.42 %.
  Intervalo de confianza del 95 %: (97.00, 99.83)
  - Sensibilidad: 60.00 %
  - Especificidad: 100.00 %

Tabla 3.41: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de Random Forest del Corpus II con el conjunto de datos A.

3.4.2.4. Conjunto de datos B

Modelo óptimo: 21 variables. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada. Número de variables seleccionadas aleatoriamente por cada partición: 17.

- Estadísticos: P90Or, MediaOr, MedianaOr, KurtosisOr, VarianzaTratadaB, KurtosisTratadaB, KurtosisTratadaN, SkewnessTratadaN.
- DCT: imagen sin binarizar: 1, 2, 6, 7, 15, 16, 29; imagen binarizada: 4, 9, 13, 15, 19, 29.

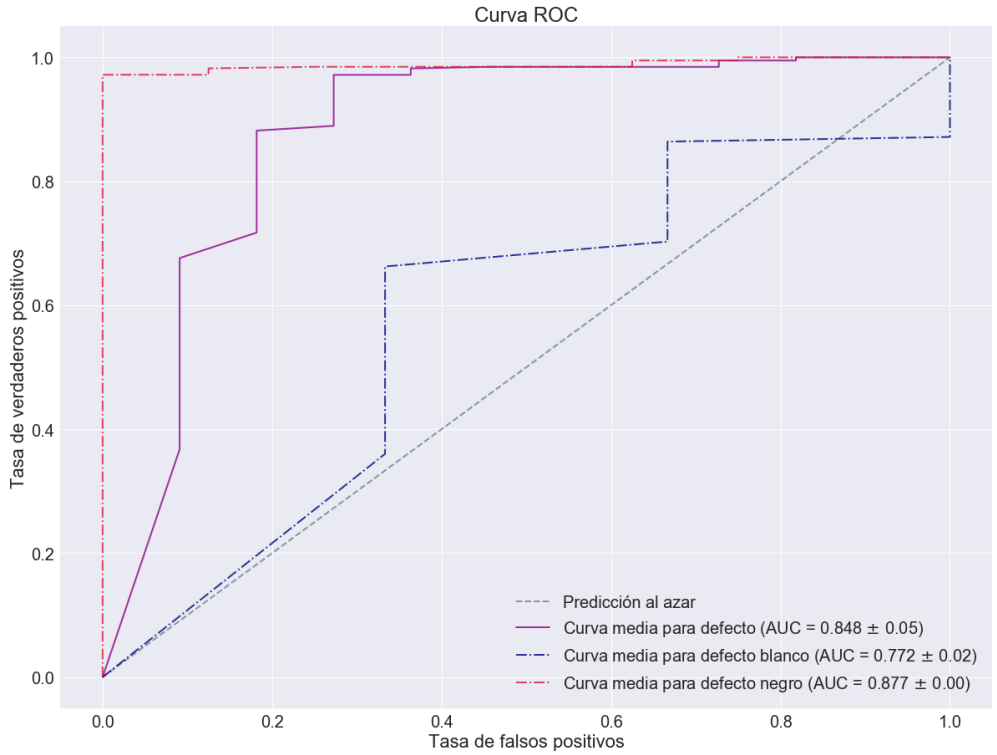

Figura 3.32: Curvas ROC de Random Forest empleando el conjunto de datos B del Corpus II.

Resultados de Random Forest en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.215 ± 0.07	0.848 ± 0.05 {	Negro: 0.877 ± 0.00 Blanco: 0.772 ± 0.02

Tabla 3.42: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo de Random Forest creado a partir del conjunto B de datos del Corpus II.

Tanto las estimaciones de las áreas bajo la curva como la EER indican que el clasificador obtenido tiene potencial para logar resultados satisfactorios (Tabla 3.42). En la Figura 3.32 vemos como la detección de defectos en general es el promedio de una excelente detección de defectos negros y una discriminación de defectos blancos peor.

En lo que se refiere al Corpus II, el resultado es muy satisfactorio (Tabla 3.43). Sin necesidad de aportar pesos a las observaciones y con tan 21 variables, parece ser superior en cuanto a precisión a muchos modelos anteriores. Aún así el el número de tableros con defectos mal clasificados es elevado; la sensibilidad no llega ni al 50 %.

Sistema Random Forest -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3876	70	3946
	Con defecto	17	37	54
		3893	107	4000

- Tasa de acierto: 97.83 %.
  Intervalo de confianza del 95 %: (96.18, 99.48)
  - Sensibilidad: 34.58 %
  - Especificidad: 99.56 %

Tabla 3.43: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo de Random Forest del Corpus II con el conjunto de datos B.

La Tabla 3.44 refleja la clara superioridad de los modelos Random Forest con respecto al sistema de referencia y a otros modelos creados anteriormente. El área bajo la curva ROC de defecto general alcanza el valor más elevado, superando al análisis discriminante y los árboles de decisión. La tasa de equierror de los modelos también es muy satisfactoria y en muchos es la más baja de todos los clasificadores desarrollados hasta el momento.

En el caso del Corpus I logramos dos clasificadores de excelente precisión y dimensionalidad muy baja. En el Corpus II se obtienen otros dos clasificadores de un número de variables similar con una precisión parecida. El resultado del B tiene menos sesgo que el del A, al no contar con observaciones añadidas artificialmente.

Resultados de los Random Forest

	Corpus I				Corpus II			
	Variables	EER	AUC	Tasa de acierto	Variables	EER	AUC	Tasa de acierto
Conjunto A	16	0.066	0.957	98.45 %	14	0.222	0.884	98.42 %
Conjunto B	36	0.021	0.996	98.80 %	21	0.215	0.848	97.83 %

Tabla 3.44: Número de variables, AUC y EER, así como tasa de acierto estimada por validación cruzada del modelo seleccionado, de los Random Forest inducidos a partir de los conjuntos de datos A y B en ambos Corpus.

3.4.3. AdaBoost

Los modelos de AdaBoost se han desarrollado tal y como aparece en la Sección 2.3.2. Con cada modelo aparece una figura que refleja la importancia de las variables en el modelo. Se han ajustado mediante validación cruzada los siguientes parámetros: número de árboles, profundidad máxima y criterio de aprendizaje de parámetros.

Corpus I

3.4.3.1. Conjunto de datos A

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 100.
  - Aprendizaje de parámetros: Zhu.
  - Máxima profundidad de los árboles: 3.
- Número de variables: 70. Tal y como vemos en la Figura 3.33, la componente principal 4 de la imagen original tiene una relevancia superior con respecto al resto de variables. Otras dos componentes principales de la imagen original y el área de error relativa en negro parecen aportar información significativa.

Existen diferencias significativas entre curvas ROC (Figura 3.34): la de defecto general es excelente pero las de defecto negro y blanco recogen menos área. La tasa de equierror es bastante baja (Tabla 3.45).

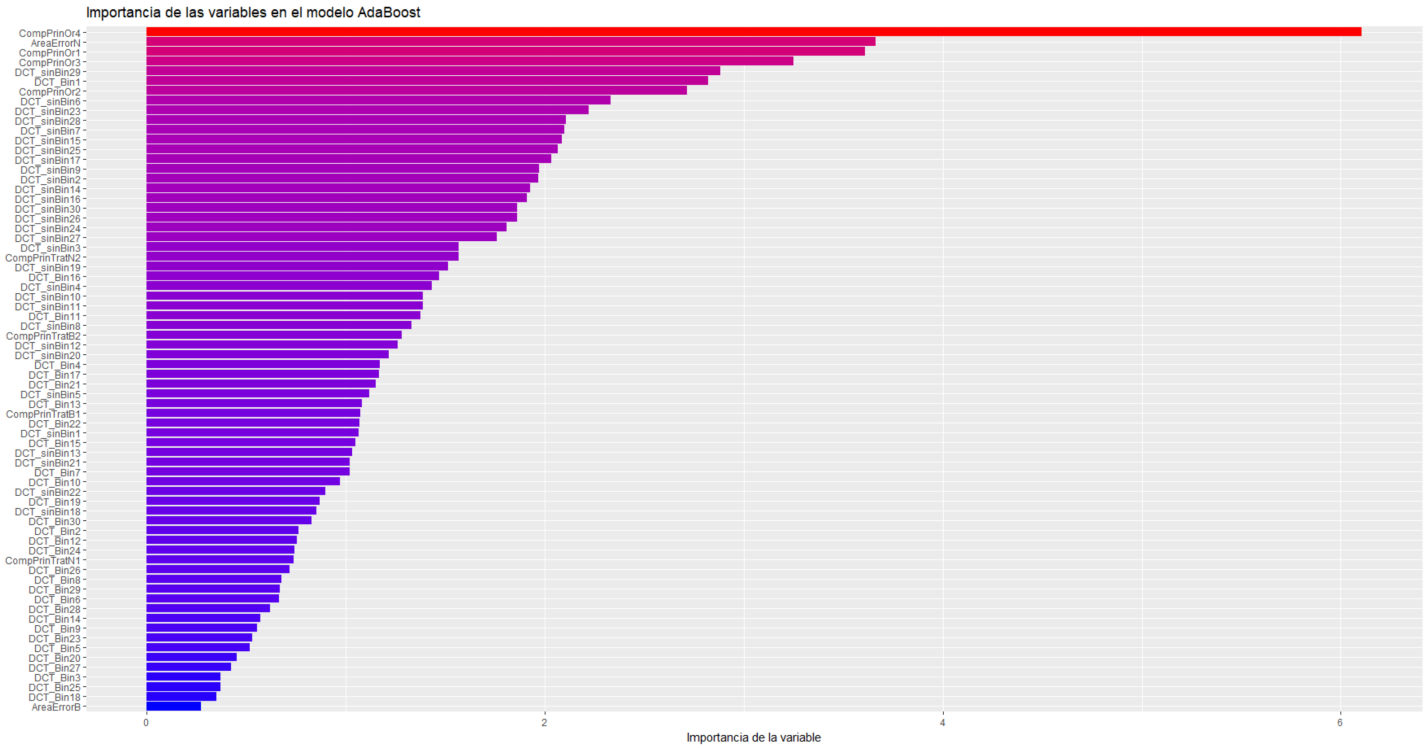

Figura 3.33: Variables según relevancia a la hora de inducir los árboles de decisión básicos del modelo AdaBoost con el conjunto de datos A (PCA) en el Corpus I.

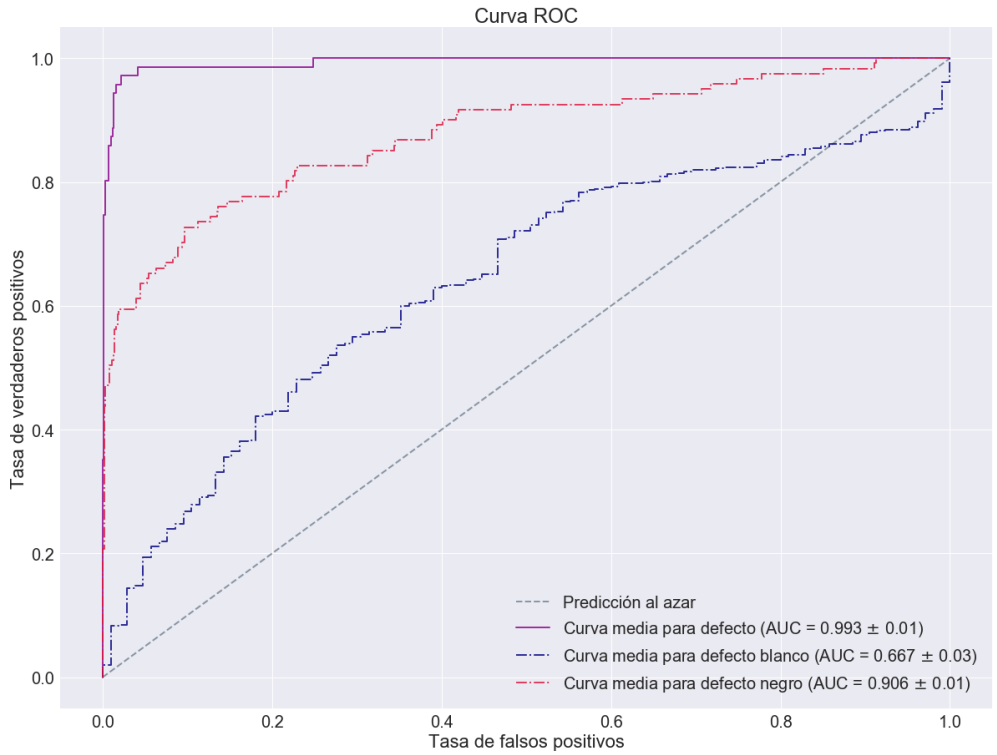

Figura 3.34: Curvas ROC del modelo AdaBoost empleando el conjunto de datos A del Corpus I.

Resultados del AdaBoost en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.027 ± 0.01	0.993 ± 0.01	Negro: 0.906 ± 0.01 Blanco: 0.667 ± 0.03

Tabla 3.45: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo AdaBoost creado a partir del conjunto A de datos del Corpus I.

El modelo AdaBoost con el conjunto de datos A del Corpus I logra una tasa de acierto altamente satisfactoria (3.46) superando la asociada al sistema de referencia y a los modelos de discriminante y árboles de decisión. La sensibilidad toma un valor alto inferior al de la especificidad.

AdaBoost -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	652	65	717
	Con defecto	57	6393	6450
		709	6458	7167

- Tasa de acierto: 98.29 %.
  Intervalo de confianza del 95 %: (97.19, 99.38)
  - Sensibilidad: 91.96 %
  - Especificidad: 98.99 %

Tabla 3.46: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo AdaBoost del Corpus I con el conjunto de datos A.

3.4.3.2. Conjunto de datos B

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 175.
  - Aprendizaje de parámetros: Zhu.
  - Máxima profundidad de los árboles: 4.
- Número de variables: 83. Tal y como vemos en la Figura 3.35, la kurtosis de la imagen tratada en negro y la media y apuntamiento de la imagen original ocupan las primeras posiciones en cuanto a importancia.

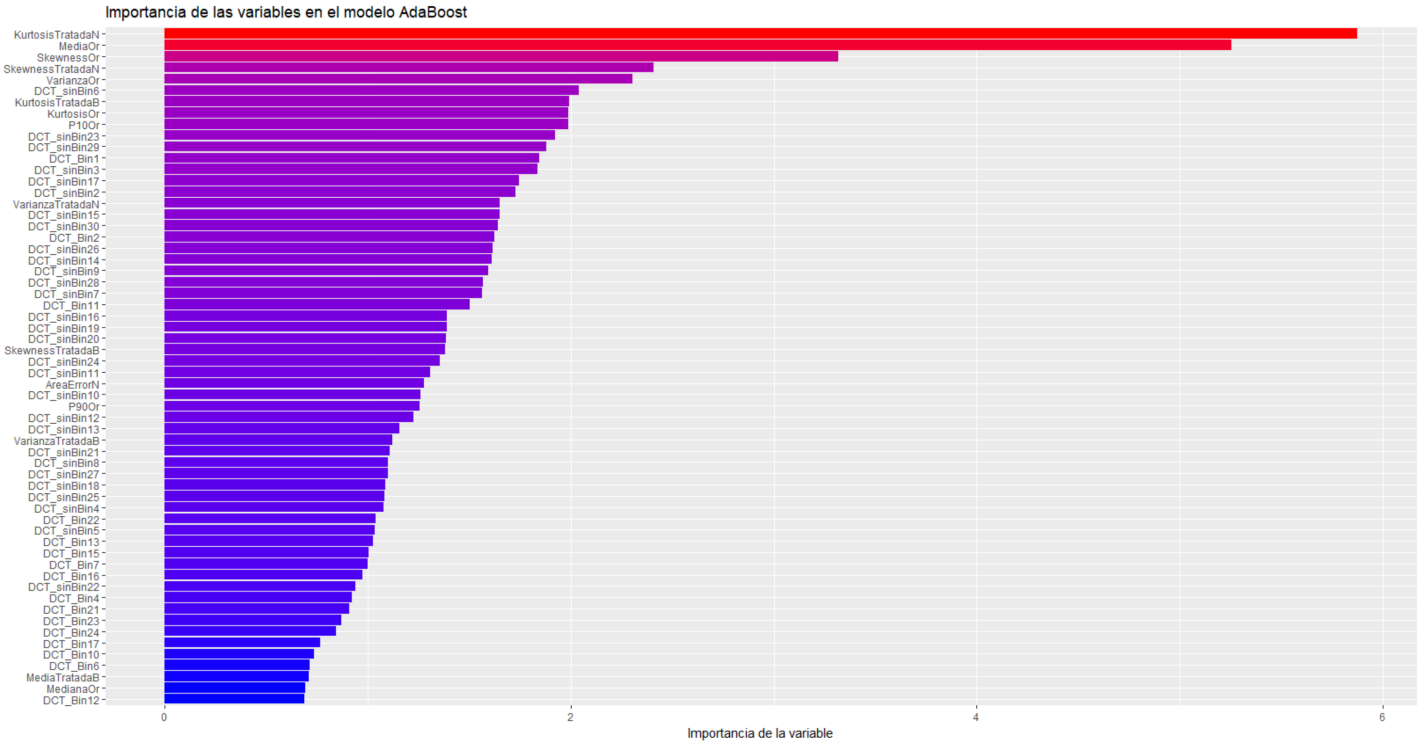

Figura 3.35: Sesenta variables de mayor relevancia a la hora de inducir los árboles de decisión básicos del modelo AdaBoost con el conjunto de datos B (Estadísticos) en el Corpus I.

Se alcanza el valor de AUC de defecto general más alto y la tasa de equierror más baja de los modelos estudiados (Tabla 3.47), empatando al modelo Random Forest inducido con este mismo conjunto de datos. Existe una clara disparidad en la capacidad discriminatoria de defectos negros y blancos (Figura 3.36).

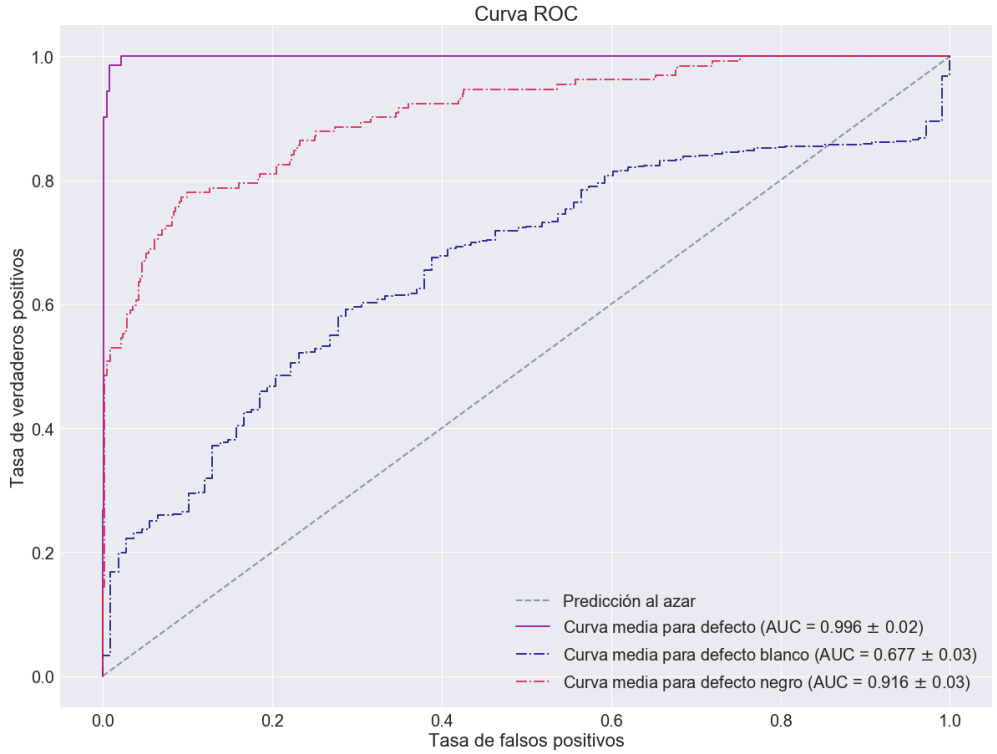

Figura 3.36: Curvas ROC del modelo AdaBoost empleando el conjunto de datos B del Corpus I.

Resultados del AdaBoost en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.021 ± 0.01	0.996 ± 0.02	Negro: 0.916 ± 0.03 Blanco: 0.677 ± 0.03

Tabla 3.47: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo AdaBoost creado a partir del conjunto B de datos del Corpus I.

AdaBoost -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	663	47	710
	Con defecto	46	6411	6457
		709	6458	7167

- Tasa de acierto: 98.70 %.
  Intervalo de confianza del 95 %: (97.74, 99.65)
  - Sensibilidad: 99.27 %
  - Especificidad: 93.51 %

Tabla 3.48: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo AdaBoost del Corpus I con el conjunto de datos B.

La tasa de acierto asociada al clasificador es la segunda más elevada (Tabla 3.48), siendo unicamente superada por una cantidad poco significativa por el clasificador Random Forest creado con este conjunto de datos. La especificidad, aun tomando un valor más pequeño que el adquirido por la sensibilidad, alcanza un resultado altamente satisfactorio.

Corpus II

3.4.3.3. Conjunto de datos A

Para la correcta inducción del clasificador, se han duplicado todas las muestras con defectos del conjunto de datos.

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 150.
  - Aprendizaje de parámetros: Freund.
  - Máxima profundidad de los árboles: 3.
- Número de variables: 70. Tal y como vemos en la Figura 3.37, el conjunto de variables relevantes es mucho mayor que en los modelos del Corpus I. El modelo tiene una fuerte dependencia de las componentes de la DCT de la imagen sin binarizar.

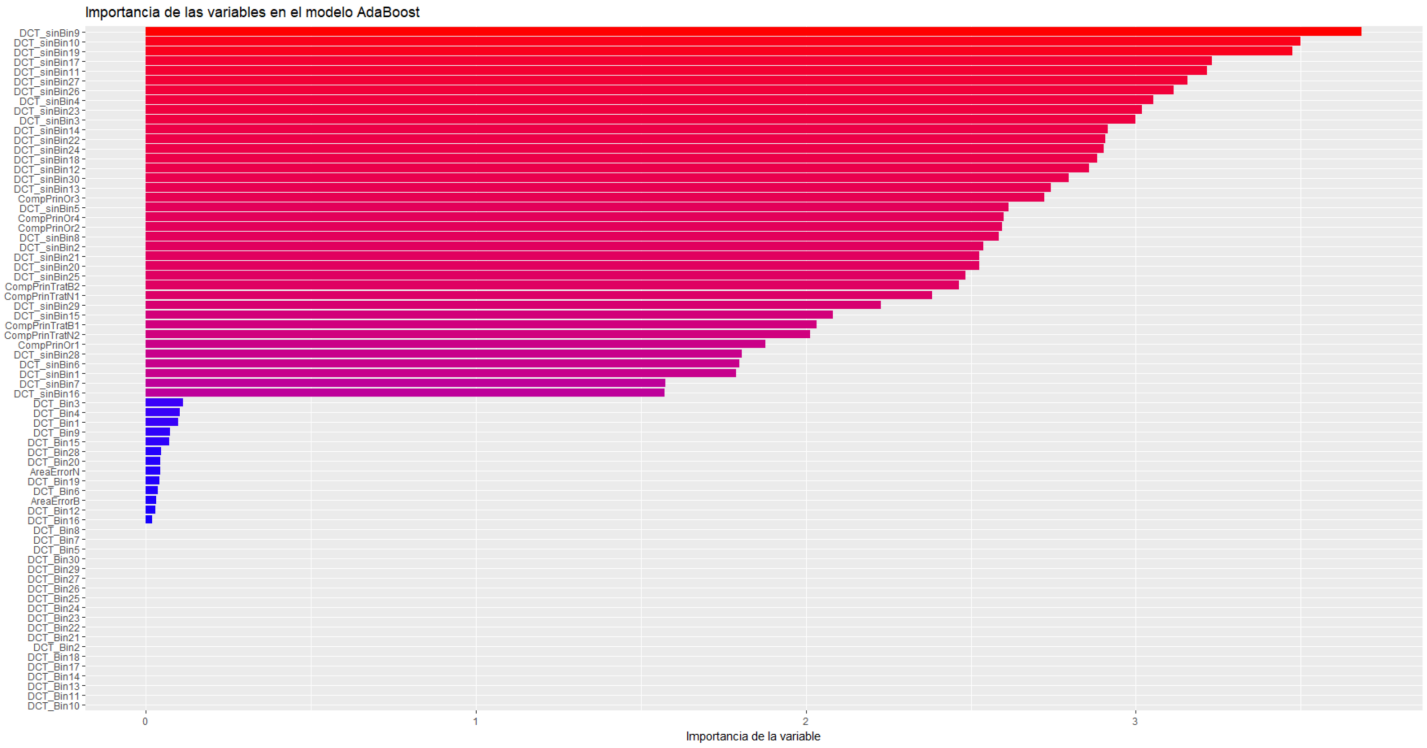

Figura 3.37: Variables según relevancia a la hora de inducir los árboles de decisión básicos del modelo AdaBoost con el conjunto de datos A (PCA) en el Corpus II.

Resultados del AdaBoost en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.142 ± 0.04	0.927 ± 0.07 {	Negro: 0.923 ± 0.08 Blanco: 0.916 ± 0.15

Tabla 3.49: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo AdaBoost creado a partir del conjunto A de datos del Corpus II.

Las curvas ROC (Figura 3.38) reflejan un claro equilbrio en el que todas ellas se pueden calificar de muy buenas. La tasa de equierror es la más baja de los modelos inducidos hasta el momento, siendo inferior al 15 % (Tabla 3.49).

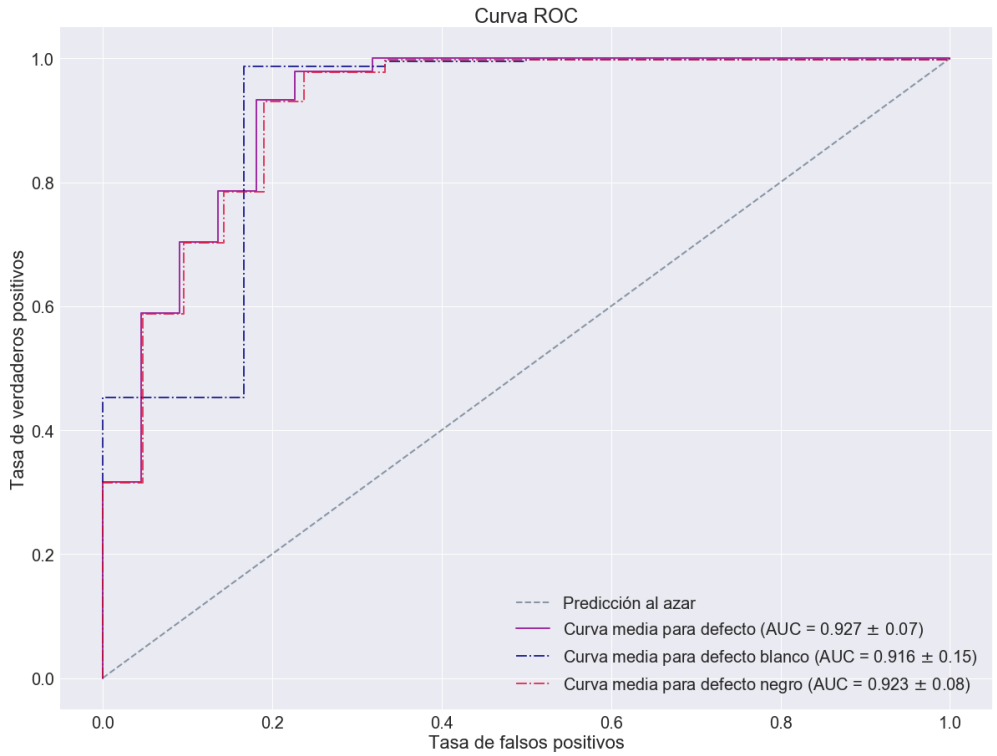

Figura 3.38: Curvas ROC del modelo AdaBoost empleando el conjunto de datos A del Corpus II.

Los resultados del modelo son los más satisfactorios del Corpus II (Figura 3.50): excelente tasa de acierto con una especificidad prácticamente perfecta y una sensibilidad alta especialmente si la comparamos con la de los modelos anteriores.

AdaBoost -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3885	41	3926
	Con defecto	8	173	181
		3893	214	4107

- Tasa de acierto: 98.80 %.
  Intervalo de confianza del 95 %: (97.73, 99.87)
  - Sensibilidad: 80.84 %
  - Especificidad: 99.79 %

Tabla 3.50: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo AdaBoost del Corpus II con el conjunto de datos A.

3.4.3.4. Conjunto de datos B

Para la correcta inducción del clasificador, se han duplicado la mitad de las muestras con defectos de los datos.

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 250.
  - Aprendizaje de parámetros: Breiman.
  - Máxima profundidad de los árboles: 9.
- Número de variables: 83. Tal y como vemos en la Figura 3.39, el conjunto de variables de importancia en el modelo es mucho más reducido que el modelo del conjunto A. La kurtosis de la imagen tratada en negro tiene una relevancia muy significativa en comparación con el resto de variables.

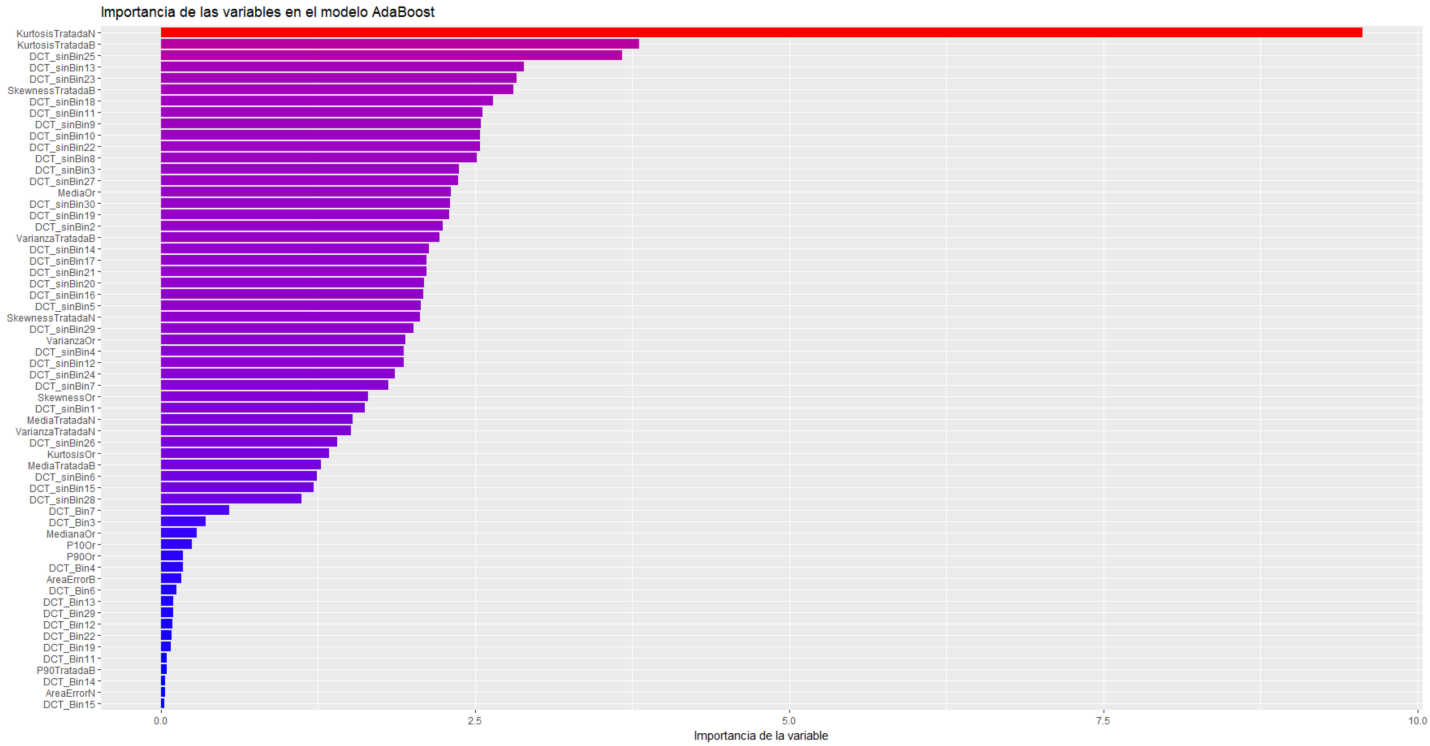

Figura 3.39: Sesenta variables de mayor relevancia a la hora de inducir los árboles de decisión básicos del modelo AdaBoost con el conjunto de datos B (Estadísticos) en el Corpus II.

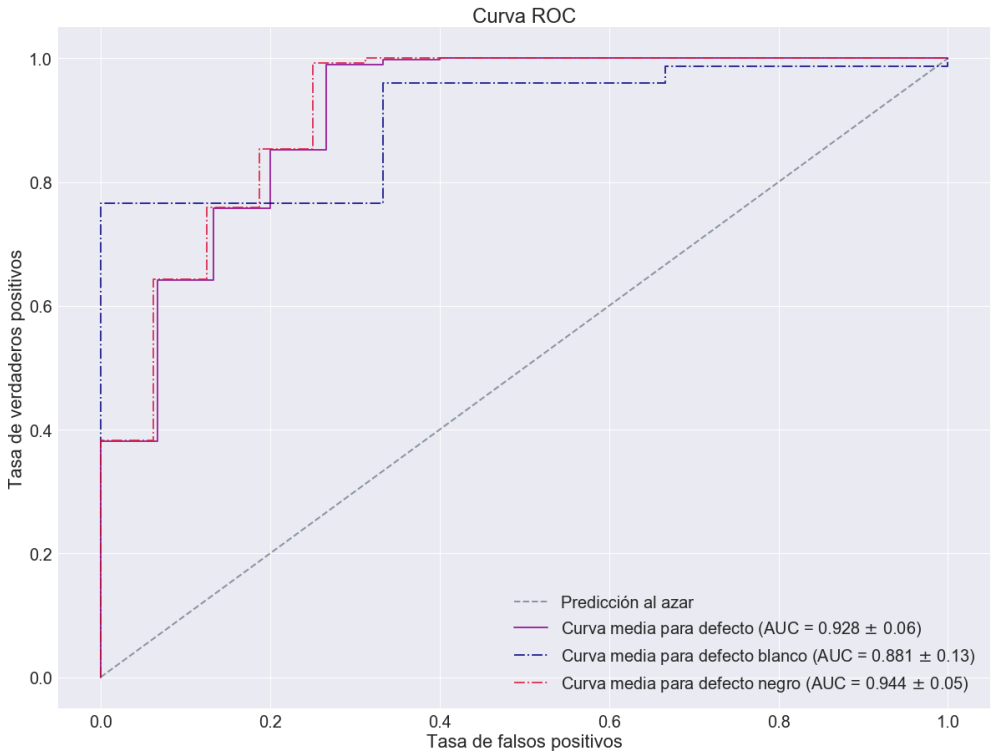

Figura 3.40: Curvas ROC del modelo AdaBoost empleando el conjunto de datos B del Corpus II.

Resultados del AdaBoost en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.149 ± 0.07	0.928 ± 0.06 {	Negro: 0.944 ± 0.05 Blanco: 0.881 ± 0.13

Tabla 3.51: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo AdaBoost creado a partir del conjunto B de datos del Corpus II.

Al igual que con el conjunto de datos A, AdaBoost logra un equilibrio muy positivo entre las tres curvas ROC (Figura 3.40). La tasa de equierror es muy satisfactoria e inferior al 15 % (Tabla 3.51).

El modelo obtenido obtiene un excelente resultado algo inferior al del conjunto A de datos (Tabla 3.52). La tasa de acierto es muy alta y la sensibilidad es la segunda más elevada obtenida entre los modelos estudiados.

AdaBoost -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3885	49	3934
	Con defecto	8	111	119
		3893	160	4053

- Tasa de acierto: 98.59 %.
  Intervalo de confianza del 95 %: (97.26, 99.92)
  - Sensibilidad: 69.37 %
  - Especificidad: 99.79 %

Tabla 3.52: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo AdaBoost del Corpus II con el conjunto de datos B.

La Tabla 3.53 muestra los resultados obtenidos a partir de los modelos AdaBoost. En el caso del Corpus I, el conjunto B empata la tasa de equierror más baja obtenida con Random Forest. Los valores de AUC y tasa de acierto son altamente similares a los de Random Forest, reflejando la clara superioridad de esta metodología y el AdaBoost con respecto a los modelos anteriores.

En el Corpus II se logran los resultados más satisfactorios del estudio, superando los valores de AUC, EER y precisión de los modelos anteriores además de poseer una sensibilidad bastante buena. Con respecto a Random Forest, los resultados de AdaBoost son algo mejores a cambio de una dimensionalidad mucho más grande.

Resultados Adaptive Boosting

	Corpus I				Corpus II			
	Variables	EER	AUC	Tasa de acierto	Variables	EER	AUC	Tasa de acierto
Conjunto A	70	0.027	0.993	98.29 %	70	0.142	0.927	98.80 %
Conjunto B	83	0.021	0.996	98.70 %	83	0.149	0.928	98.59 %

Tabla 3.53: Número de variables, AUC y EER, así como tasa de acierto estimada por validación cruzada del modelo seleccionado, de los modelos AdaBoost a partir de los conjuntos de datos A y B en ambos Corpus.

3.4.4. Bayesian Additive Regression Trees

A lo largo de esta sección se exponen los resultados obtenidos con los modelos BART creados así como los parámetros seleccionados como óptimos. En todos ellos se han empleado la totalidad de variables de los conjuntos y se ha hecho un estudio de importancia relativa en forma de gráficos, tal y como se explico en la sección 2.3.3.

Corpus I

3.4.4.1. Conjunto de datos A

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 50.
  - Parámetros de poda: $\alpha = 0.99, \beta = 2$.
- Número de variables: 70. Tal y como vemos en la Figura 3.41, la componente principal 4 de la imagen original destaca sobre el resto de variable. El área de error relativa en negro así como diversas componentes de la DCT de la imagen binarizada y sin binarizar completan otras posiciones elevadas en el ranking de importancia.

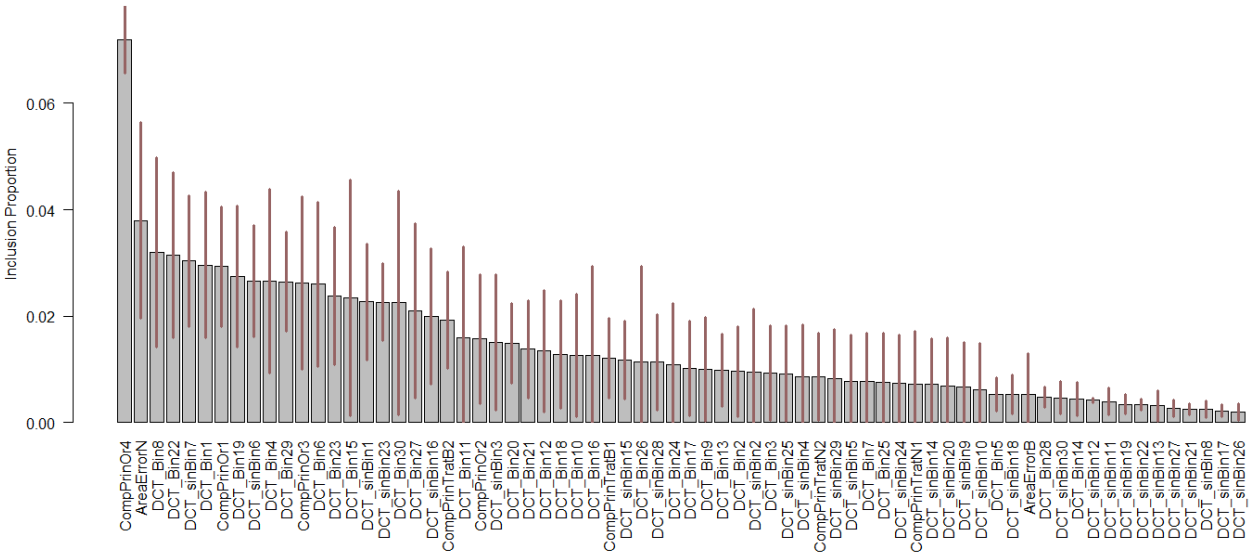

Figura 3.41: Frecuencia de inclusión de las variables con el conjunto de variables A (PCA) a la hora de inducir los árboles de decisión básicos del modelo BART en el Corpus I.

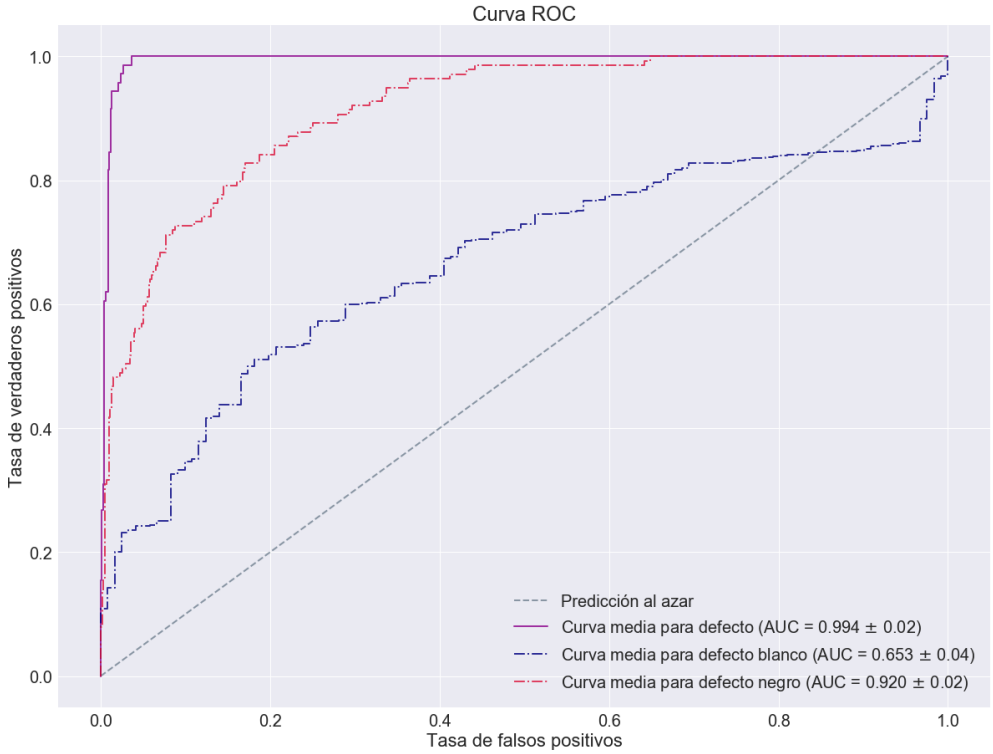

Figura 3.42: Curvas ROC de BART empleando el conjunto de datos A del Corpus I.

Resultados de BART en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.027 ± 0.01	0.994 ± 0.02	Negro: 0.920 ± 0.02 Blanco: 0.663 ± 0.04

Tabla 3.54: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo BART creado a partir del conjunto A de datos del Corpus I.

Las curvas ROC asociadas al clasificador (Figura 3.42) indican que su capacidad discriminatoria es bastante buena, obteniendo resultados de un calibre similar a los previos de Random Forest o AdaBoost. La EER tiene un valor bastante bajo (Tabla 3.54). Las AUC indican cierto desequilibrio en la detección de defectos según el tipo.

El modelo obtenido es muy equilibrado. La detección de tableros sin defectos sigue siendo algo peor (Tabla 3.55), haciendo que la especificidad sea más baja. La precisión es superior al sistema de referencia, discriminante y los árboles de decisión creados y algo peor que la obtenida con los modelos Random Forest y AdaBoost.

Bayesian Additive Regression Trees -Conjunto de datos A, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	643	65	708
	Con defecto	66	6393	6459
		709	6458	7167

- Tasa de acierto: 98.17 %.
  Intervalo de confianza del 95 %: (97.04, 99.30)
  - Sensibilidad: 98.99 %
  - Especificidad: 90.69 %

Tabla 3.55: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo BART del Corpus I con el conjunto de datos A.

3.4.4.2. Conjunto de datos B

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 50.
  - Parámetros de poda: $\alpha = 0.99, \beta = 3$
- Número de variables: 83. En la Figura 3.43 de relevancia de variables destacan estadísticos de la imagen original (media, apuntamiento, mediana), de la imagen tratada en negro (kurtosis, apuntamiento) y de la imagen tratada en blanco (mediana) así como componentes DCT de la imagen binarizada. Parece que la importancia de las variables es más equilibrada que en el caso del modelo del conjunto de datos A.

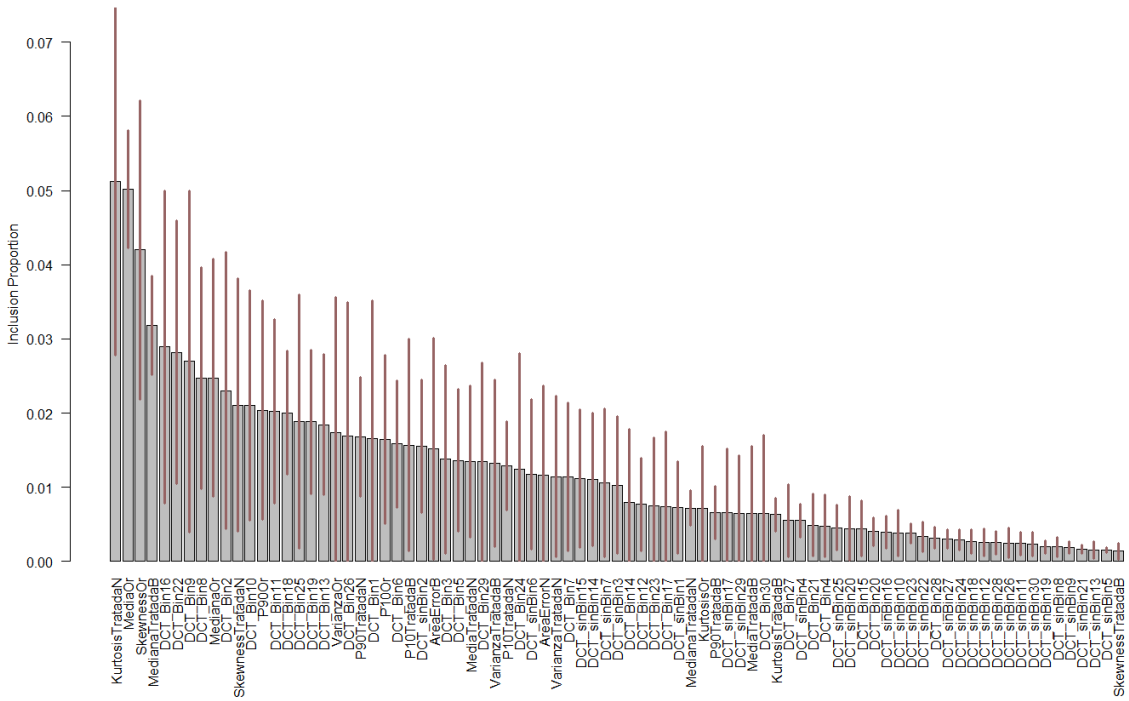

Figura 3.43: Frecuencia de inclusión de las variables con el conjunto de variables B (EST) a la hora de inducir los árboles de decisión básicos del modelo BART en el Corpus I.

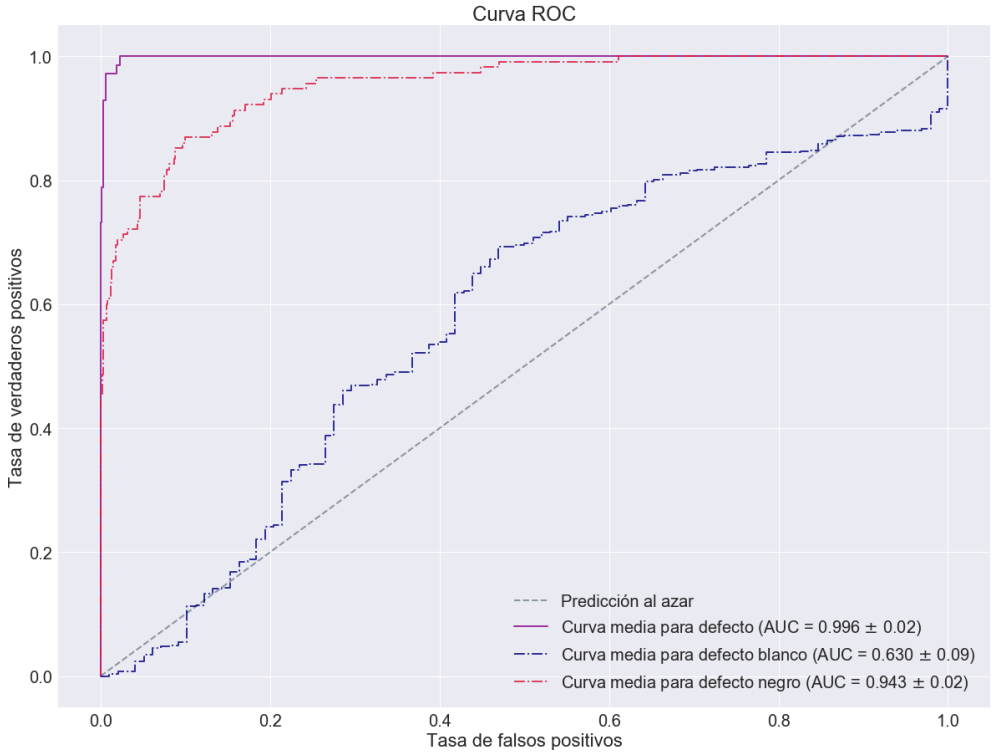

Figura 3.44: Curvas ROC de BART empleando el conjunto de datos B del Corpus I.

Resultados de BART en el Corpus I			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.025 ± 0.01	0.996 ± 0.02 {	Negro: 0.943 ± 0.02 Blanco: 0.630 ± 0.09

Tabla 3.56: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo BART creado a partir del conjunto B de datos del Corpus I.

La curva ROC (Figura 3.44) y los valores de la tasa de equierror (Tabla 3.56) reflejan como el modelo tiene un funcionamiento excelente en lo que se refiere a la detección en general. Estratificando por el tipo de defecto, parece que la detección de defectos blancos es mucho peor que la de negros como pasaba con el clasificador del conjunto A.

El modelo obtenido tiene una precisión inferior a la del modelo del conjunto A (Tabla 3.57), aunque sigue siendo superior a la del sistema de referencia y los modelos de discriminante. Aunque la detección de tableros con defectos es bastante buena, los tableros sin defectos suponen una cierta dificultad para el clasificador.

Bayesian Additive Regression Trees -Conjunto de datos B, Corpus I-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	581	44	625
	Con defecto	128	6414	6542
		709	6458	7167

- Tasa de acierto: 97.60 %.
  Intervalo de confianza del 95 %: (96.31, 98.89)
  - Sensibilidad: 99.31 %
  - Especificidad: 81.95 %

Tabla 3.57: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo BART del Corpus I con el conjunto de datos B.

Corpus II

3.4.4.3. Conjunto de datos A

Se han re-muestreado el doble de veces las observaciones de imágenes con defectos ya que conforman un porcentaje relativamente bajo del conjunto de datos. De no hacerse así se obtenía un clasificador demasiado desbalanceado.

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 50.
  - Parámetros de poda: $\alpha = 0.99, \beta = 1$
- Número de variables: 70. La Figura 3.45 nos destaca de forma similar a la correspondiente en el Corpus I la importancia de las componentes principales extraídas de la imagen original. También diversas componentes de la DCT de la imagen sin binarizar aparecen en las primeras posiciones. La importancia de las variables es relativamente equilibrada.

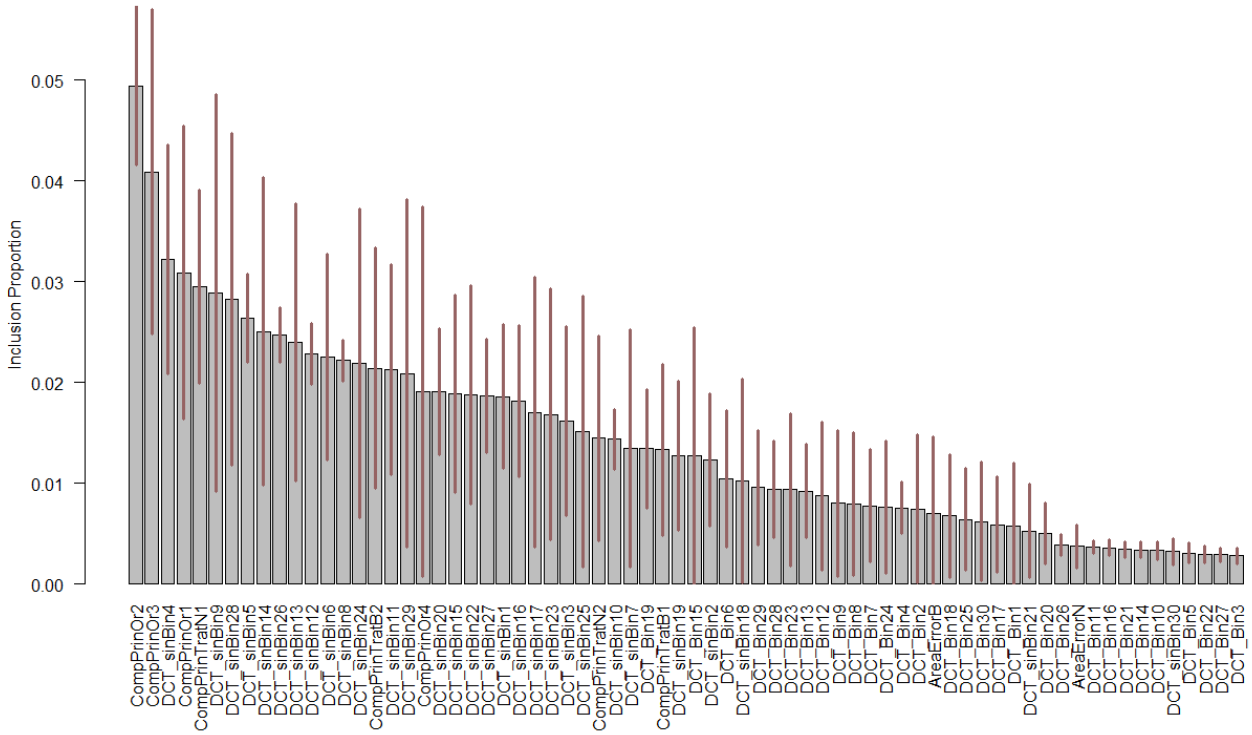

Figura 3.45: Frecuencia de inclusión de las variables con el conjunto de variables A (PCA) a la hora de inducir los árboles de decisión básicos del modelo BART en el Corpus II.

Resultados de BART en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos A	0.156 ± 0.04	0.925 ± 0.11 {	Negro: 0.918 ± 0.05 Blanco: 0.869 ± 0.14

Tabla 3.58: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo BART creado a partir del conjunto A de datos del Corpus II.

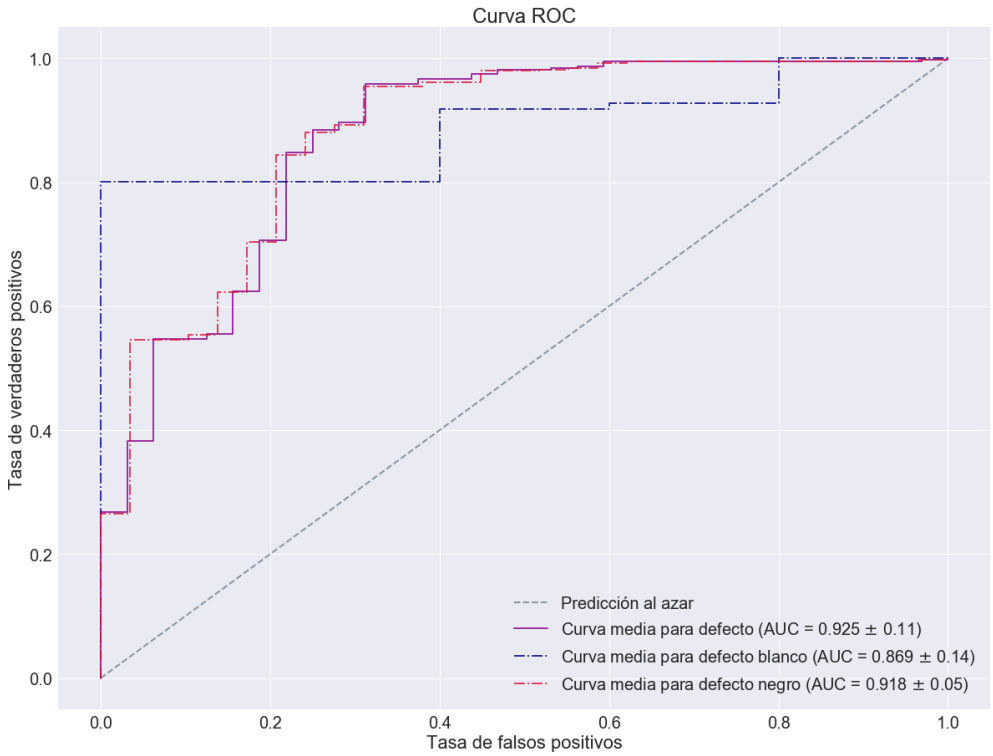

Figura 3.46: Curvas ROC de BART empleando el conjunto de datos A del Corpus II.

El valor del área bajo la curva ROC de defecto (Figura 3.46) así como la tasa de equierror (Tabla 3.58) indican que el funcionamiento de la detección de defectos es bastante bueno para tratarse de las imágenes de menor calidad del Corpus II. La detección de defectos negros y blancos es bastante equilibrada. Cabe recordar como estas medidas cuenta con cierto sesgo debido al uso repetido de muestras.

La Tabla 3.59 muestra como el clasificador creado no da lugar a un resultado destacable: la tasa de acierto de los modelos de discriminante lineal era mejor. La sensibilidad es excesivamente baja. El haber duplicado las muestras de observaciones con defectos para obtener estos resultados es un indicativo más de que el clasificador resultante no tiene un funcionamiento adecuado.

Bayesian Additive Regression Trees -Conjunto de datos A, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3855	229	4084
	Con defecto	38	90	128
		3893	319	4212

- Tasa de acierto: 93.66 %.
  Intervalo de confianza del 95 %: (90.90, 96.42)
  - Sensibilidad: 28.21 %
  - Especificidad: 99.02 %

Tabla 3.59: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo BART del Corpus II con el conjunto de datos A.

3.4.4.4. Conjunto de datos B

Para evitar la creación de un modelo desequilibrado, la mitad de las muestras con defectos del Corpus se han duplicado para entrenar al modelo.

Modelo óptimo. Seleccionado bajo el criterio de máxima tasa de acierto estimada por validación cruzada.

- Parámetros:
  - Número total de árboles: 50.
  - Parámetros de poda: $\alpha = 0.9, \beta = 1$

- Número de variables: 83. El ranking de importancia de las variables representado en la Figura 3.47 emplaza en las primeras posiciones estadísticos de la imagen original (media, mediana) y de la imagen tratada en negro (kurtosis, apuntamiento), de forma similar a lo ocurrido con el conjunto B del Corpus I. La importancia de la kurtosis de la imagen tratada en negro destaca sobre el resto de variables.

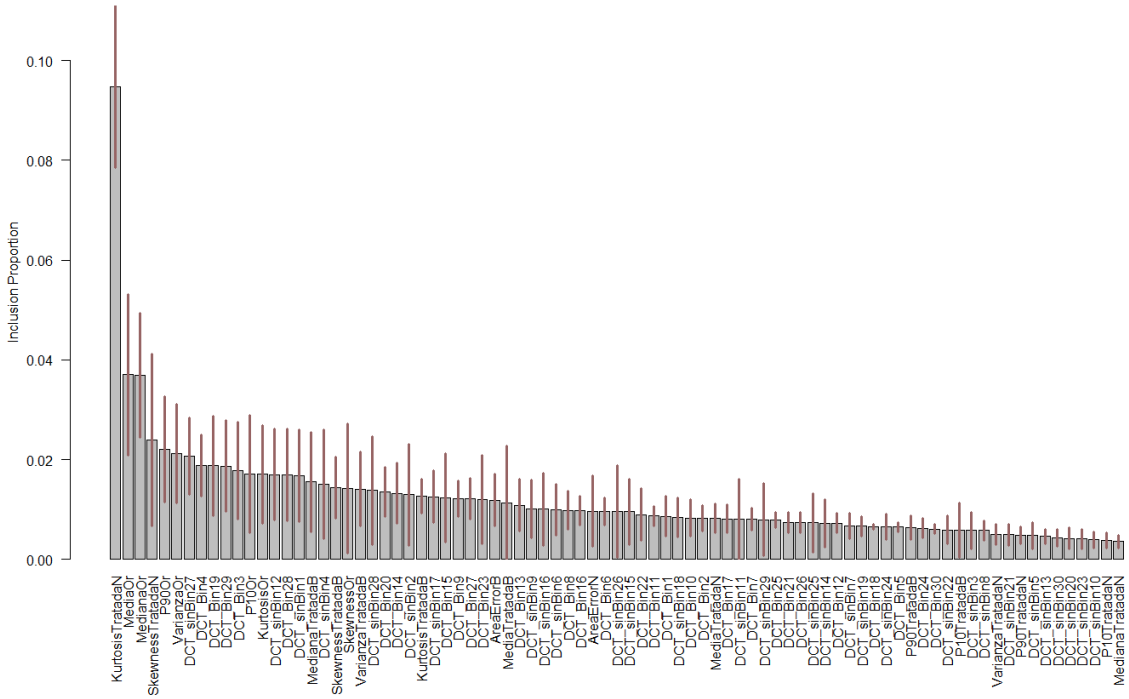

Figura 3.47: Frecuencia de inclusión de las variables con el conjunto de variables B (EST) a la hora de inducir los árboles de decisión básicos del modelo BART en el Corpus II.

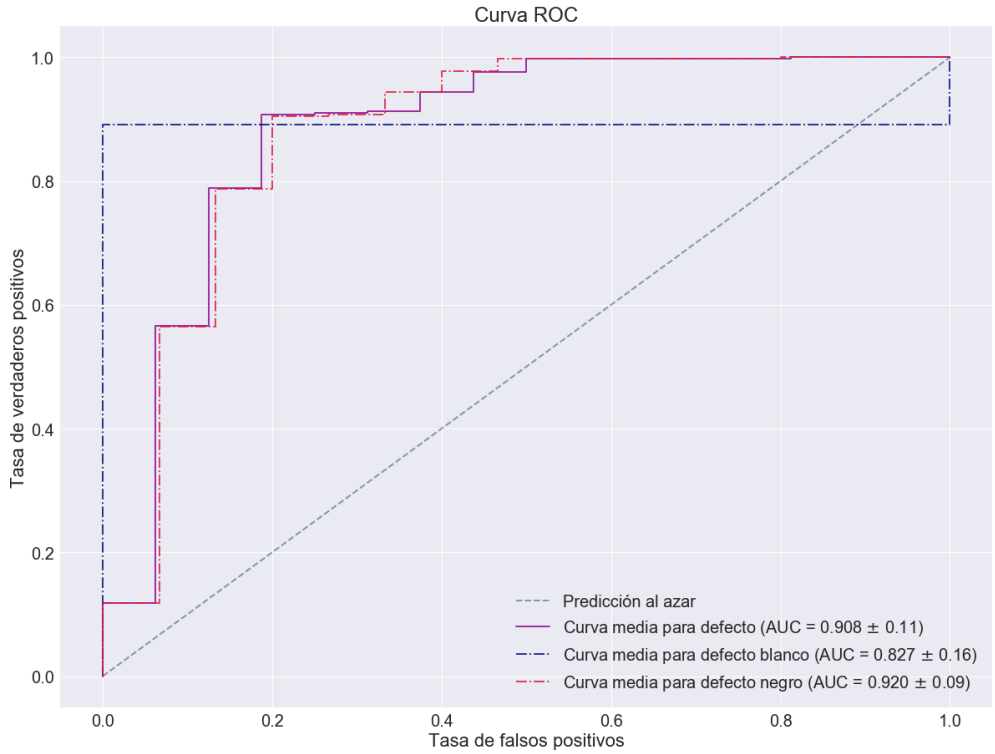

Figura 3.48: Curvas ROC de BART empleando el conjunto de datos B del Corpus II.

Las curvas ROC (Figura 3.48) parecen indicar la existencia de un equilibrio correcto en la detección de los diferentes defectos. La tasa de equierror y los valores de AUC (Tabla 3.60) indican que puede tener una buena capacidad clasificadora.

Resultados de BART en el Corpus II			
	Tasa de equierror (EER)	Area Under Curve (AUC)	
Conjunto de datos B	0.173 ± 0.06	0.908 ± 0.11 {	Negro: 0.920 ± 0.09 Blanco: 0.827 ± 0.16

Tabla 3.60: Estimación de la tasa de equierror y el área bajo la curva ROC del modelo BART creado a partir del conjunto B de datos del Corpus II.

La Tabla 3.61 muestra como la precisión del clasificador es bastante buena. La detección de tableros sin defectos es muy buena, pero los tableros con características defectuosas no son detectados correctamente en muchas ocasiones. Aún así, la sensibilidad se aproxima al 50 %.

Bayesian Additive Regression Trees -Conjunto de datos B, Corpus II-

		Clase		
		Sin Defecto	Con defecto	
Etiqueta	Sin Defecto	3873	93	3966
	Con defecto	20	67	87
		3893	160	4053

- Tasa de acierto: 97.21 %.
  Intervalo de confianza del 95 %: (95.35, 99.07)
  - Sensibilidad: 41.87 %
  - Especificidad: 99.49 %

Tabla 3.61: Matriz de confusión y tasa de acierto correspondiente del modelo óptimo BART del Corpus II con el conjunto de datos B.

La Tabla resumen 3.62 expone como los modelos BART correctamente creados resultan en tasas de acierto elevadas, mejores que las obtenidas con la discriminación y el sistema de referencia, pero no tan altas como las obtenidas con Random Forest o Adaboost.

Los valores estimados de las tasas de equierror y áreas bajo la curva son bastante satisfactorias, tanto en el Corpus I como en el II.

La dependencia de estos modelos de técnicas de computación intensiva (algoritmo de Metropolis-Hastings, muestro de Gibbs), su alta dimensionalidad y el resultado del BART creado con el conjunto B de datos del Corpus II parecen indicarnos que los resultados no son mejores por no contar con más observaciones que equilibrasen la distribución de clases. Crear una submuestra de menor tamaño equilibrada da lugar a un problema de falta de observaciones, por lo que no es una solución apta.

Resultados Bayesian Additive Regression Trees								
	Corpus I				Corpus II			
	Variables	EER	AUC	Tasa de acierto	Variables	EER	AUC	Tasa de acierto
Conjunto A	70	0.027	0.994	98.17 %	70	0.156	0.925	93.66 %
Conjunto B	83	0.025	0.996	97.60 %	83	0.173	0.908	97.21 %

Tabla 3.62: Número de variables, AUC y EER, así como tasa de acierto estimada por validación cruzada del modelo seleccionado, de los modelos Bayesian Aditive Regression Trees a partir de los conjuntos de datos A y B en ambos Corpus.

Capítulo 4

Discusión general de resultados

Este capítulo consistirá en la discusión global de los resultados alcanzados con los diferentes modelos desarrollados a lo largo de la memoria de estadística. Estos modelos se compararán con el sistema de referencia que emplea unicamente la suma de las variables del área de error relativa. Esto se debe a que sus resultados que aquellos logrados con un sistema que emplee las dos variables por separado. En la correspondiente sección de la memoria de Trabajo de Fin de Grado de ingeniería informática [1] se comentan los resultados relativos a los sistemas desarrollados en la misma (regresión logística, Support Vector Machine y redes neuronales).

Como se ha detallado anteriormente, se realizará una comparativa global con los resultados estimados (área bajo la curva ROC, tasa de equierror, tasa de acierto, sensibilidad, especificidad y F-Score) por validación cruzada de 10 particiones para los dos conjuntos de datos extraídos (Conjunto A -Componentes Principales- y Conjunto B -Estadísticos-) de los Corpus I y II de imágenes.

4.1. Comparativa de los modelos según AUC y EER

Compendio general de resultados								
			Corpus I			Corpus II		
			Variables	EER	AUC	Variables	EER	AUC
	00	Sist. Ref.	1	0.046	0.962	1	0.551	0.459
04 An.Discr.	LDA	Conjunto A	35	0.103	0.934	35	0.422	0.611
		Conjunto B	41	0.100	0.944	41	0.254	0.835
	QDA	Conjunto A	35	0.046	0.975	15	0.314	0.664
		Conjunto B	41	0.077	0.961	14	0.412	0.589
05	C4.5	Conjunto A	20	0.090	0.940	31	0.488	0.596
		Conjunto B	16	0.076	0.946	38	0.196	0.796
06	RF	Conjunto A	16	0.066	0.957	14	0.222	0.884
		Conjunto B	36	**0.021**	**0.996**	21	0.215	0.848
07	ADA	Conjunto A	70	0.027	0.993	70	**0.142**	0.927
		Conjunto B	83	**0.021**	**0.996**	83	0.149	**0.928**
08	BART	Conjunto A	70	0.027	0.994	70	0.156	0.925
		Conjunto B	83	0.025	**0.996**	83	0.173	0.908

Tabla 4.1: Número de variables, AUC y EER de los modelos estudiados en la memoria de estadística.

En la Tabla 4.1 aparecen los valores de las tasas de equierror y áreas bajo la curva ROC de defecto general de los modelos estudiados en la memoria de estadística. Se puede afirmar de forma general que los modelos arbóreos (C4.5, Random Forest, AdaBoost, BART) y QDA han resultado en su mayoría ser superiores si tomamos como criterio de comparación el AUC y el EER.

En el Corpus I una selección de variables de los datos del conjunto B con el clasificador Random Forest parece alzarse con el mejor resultado en cuanto a AUC y EER. El clasificador Adaboost logra igualar estos resultados a cambio del uso del conjunto de variables B al completo. Ocurre de forma similar con el modelo BART que, aún teniendo una tasa de equierror no significativamente mayor, logra también el mayor valor para el área bajo la curva ROC. Estos clasificadores con el conjunto de variables A obtienen medidas de la EER y AUC algo peores con la ventaja del uso de un número inferior de variables. El caso de Random Forest con este conjunto de datos es especialmente llamativo al obtener un excelente resultado con pocas variables. Por detrás de los sistemas de clasificación mentados se hallarían el QDA con 35 variables del conjunto A y, curiosamente, el sistema de referencia.

En el caso del Corpus II, los dos modelos AdaBoost logran el mejor resultado en AUC y EER respectivamente. No parece existir una diferencia significativa entre ellos exceptuando un menor número de variables al usar el conjunto de datos A. Los dos modelo BART, inducidos con los conjuntos de datos A y B, logran el siguiente juego de mejores resultados seguido de los modelos de árbol de decisión C4.5 y discriminante lineal creados con el conjunto de variables de estadísticos. Los modelos de regresión logística, SVM de núcleo RBF y los MLP logran resultados medios con tasas de equierror inferiores al 45 %. El resto de modelos logran resultados bastante pobres.

Las Figuras 4.1 y 4.2 muestran mediante gráficos de líneas las estimaciones realizadas de área bajo la curva ROC y la probabilidad complementaria a la tasa de equierror: $1 - EER$, a la que he denominado tasa de equiacierto, asociadas a los modelos creados con los conjuntos de datos A y B extraídos de los Corpus I y II respectivamente. Es importante destacar la diferencia de escala existente en ambas Figuras.

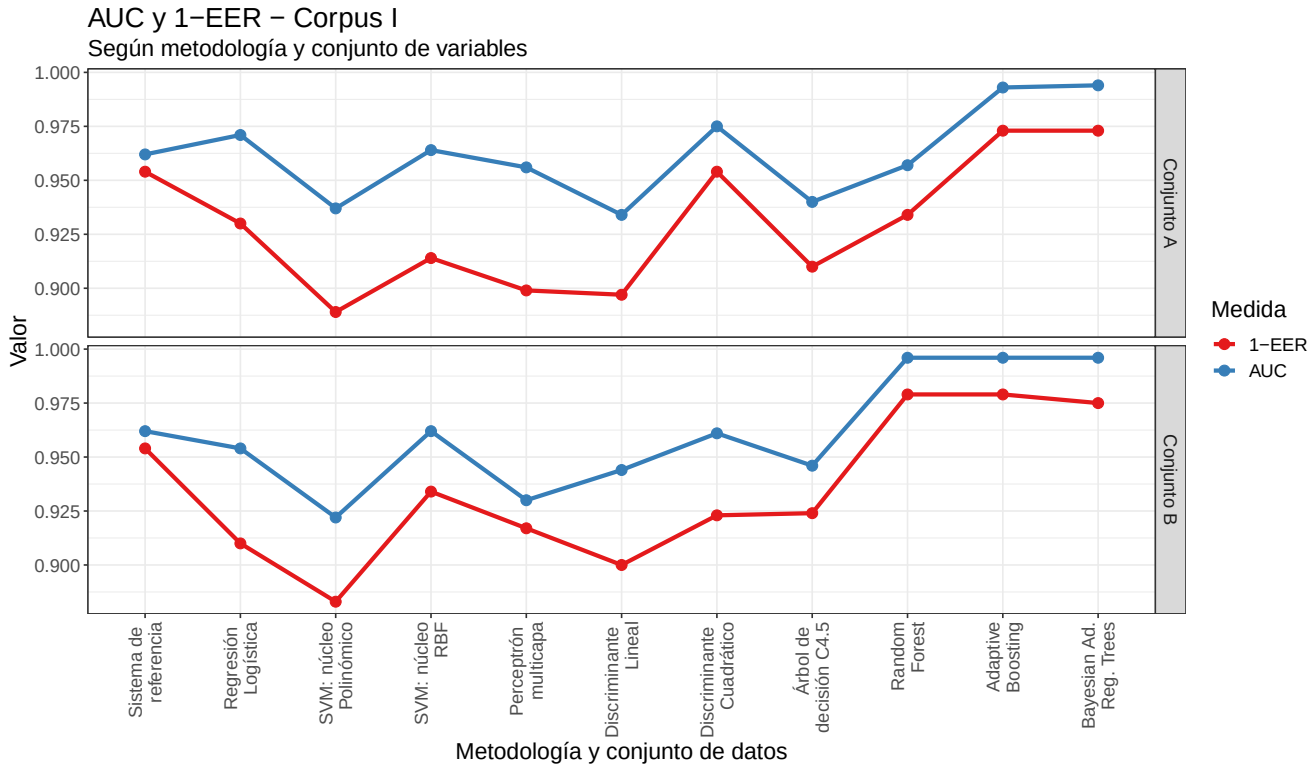

Figura 4.1: AUC y tasa de equiacierto para cada uno de los conjuntos de datos del Corpus I y metodología empleada.

Las conclusiones que podemos sacar son similares a las anteriores, destacando una mayor uniformidad de resultados en el Corpus I contrapuesta a la variedad de resultados del Corpus II fruto de la peor calidad de imágenes. Podemos ver cómo las medidas del área bajo la curva ROC y la tasa de equiacierto, y por tanto de la tasa de equierror, son bastante parejas en lo que se refiere a la calidad de las metodologías. La Figura 4.3 muestra en un solo gráfico de líneas los resultados de área bajo la curva ROC y tasa de equiacierto para ambos Corpus con cada uno de los algoritmos empleados. Los resultados de los clasificadores del Corpus I son superiores a los del Corpus II. Esto se debe, como ya se ha comentado, a la peor calidad del segundo conjunto de imágenes.

Las Figuras 4.4 y 4.5 son una visualización de la relación existente entre el AUC y el número de variables en función del conjunto de datos y la metodología empleadas en el Corpus I y Corpus II. Aunque las figuras de los dos Corpus son altamente similares, existe una diferencia de escala significativa en el eje vertical. Se podría hacer otro par de figuras con la tasa de equierror pero las representaciones que obtendríamos serían análogas a las del AUC.

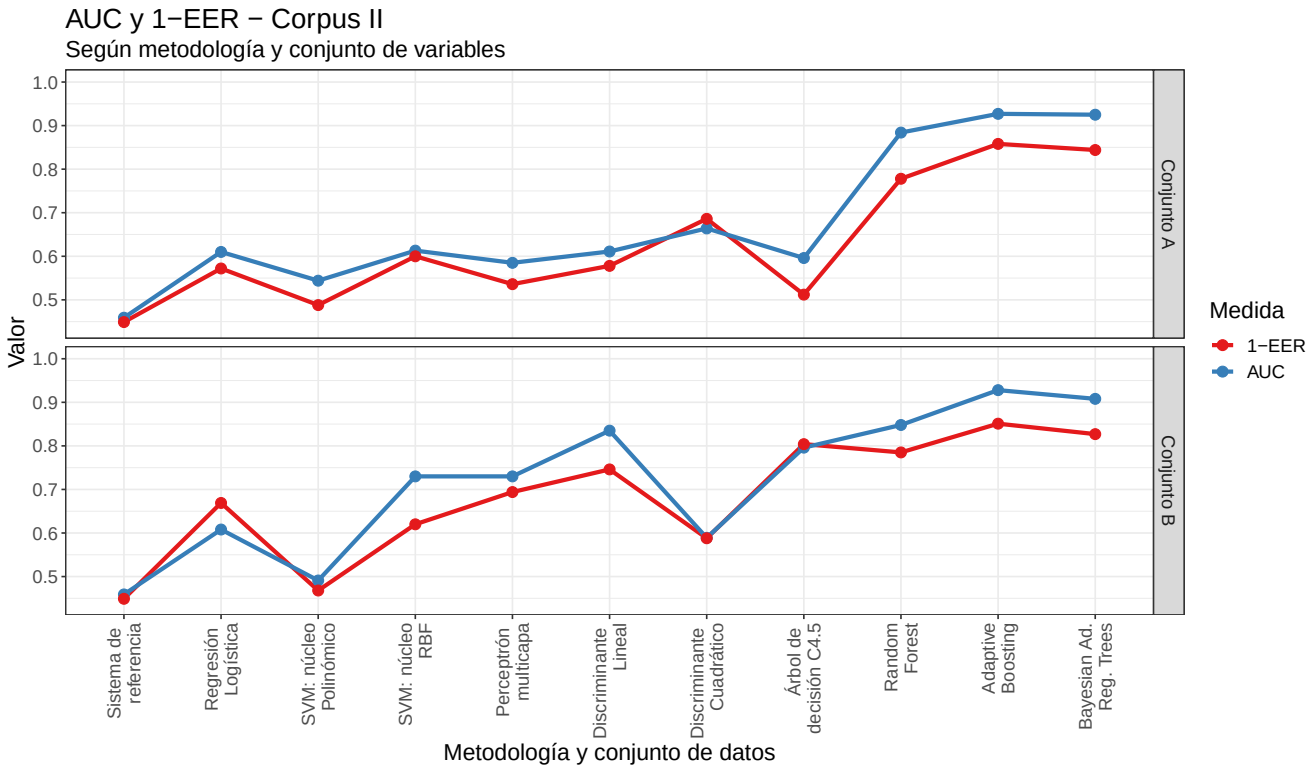

Figura 4.2: Área bajo la curva ROC y tasa de equiacierto para cada uno de los conjuntos de datos del Corpus II y metodología empleada.

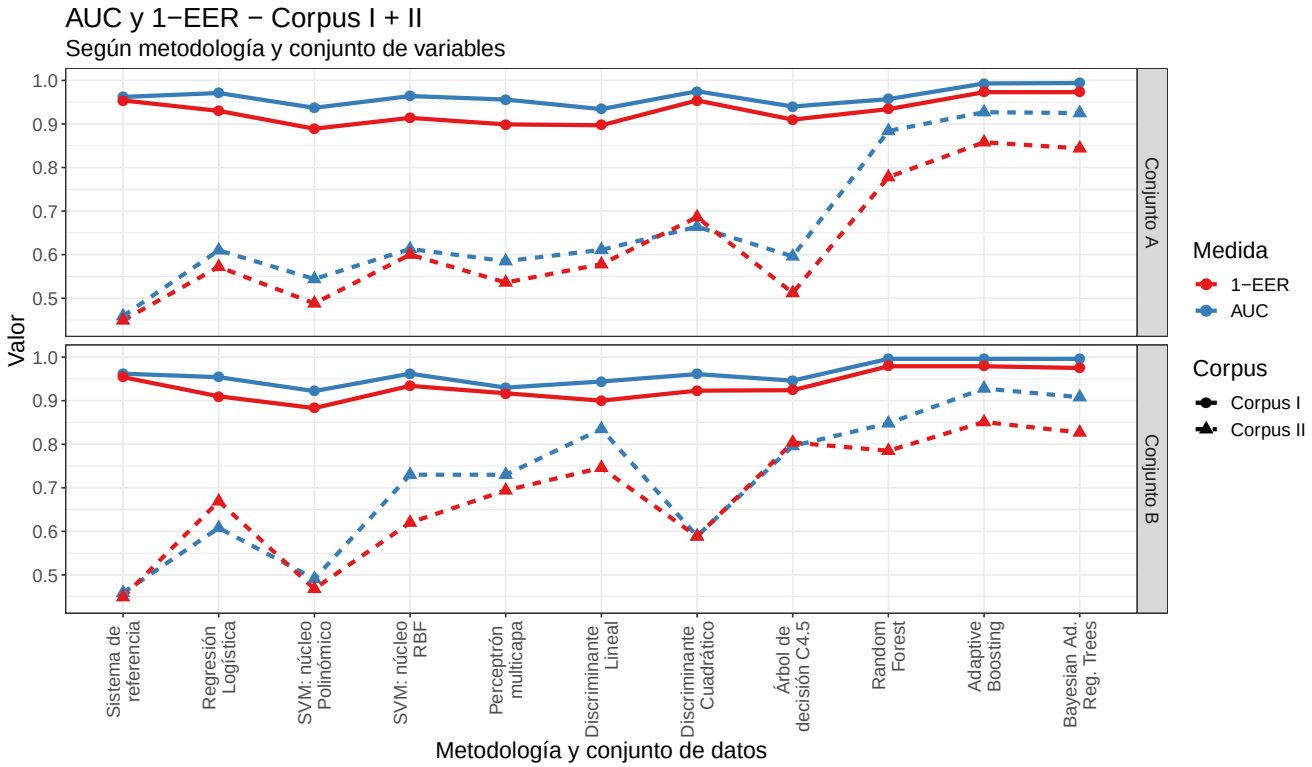

Figura 4.3: Área bajo la curva ROC y tasa de equiacierto para cada uno de los conjuntos de datos de los Corpus I y II (representados con estilos de línea diferentes) y metodología empleada.

En el Corpus I, los modelos, tal y como hemos visto, logran valores de AUC bastante altos. Los modelos de agregación de árboles de decisión mediante Boosting (AdaBoost, BART) logran resultados satisfactorios a cambio de una dimensionalidad elevada. El modelo Random Forest con el conjunto B logra un resultado igualmente satisfactorio con menos de 40 variables. Quedan por debajo los modelos de discriminación cuadrática y el simple sistema de referencia con el área de error relativa.

Los valores de área bajo la curva ROC asociados a los modelos del Corpus II tienen un rango de valores más amplio que en el Corpus I. Podemos ver cómo el sistema de referencia es insuficiente al hallarse cerca de la asignación aleatoria, posiblemente por la menor eficacia del filtrado al contar con menos resolución las imágenes del Corpus II. Los modelos de Boosting son superiores al resto de modelos con el uso del número máximo de variables disponibles en cada juego de variables. Por detrás y con una dimensionalidad inferior se encuentran los modelos de Random Forest y los modelos de discriminante lineal y árboles de decisión C4.5 inducidos con el conjunto de variables B.

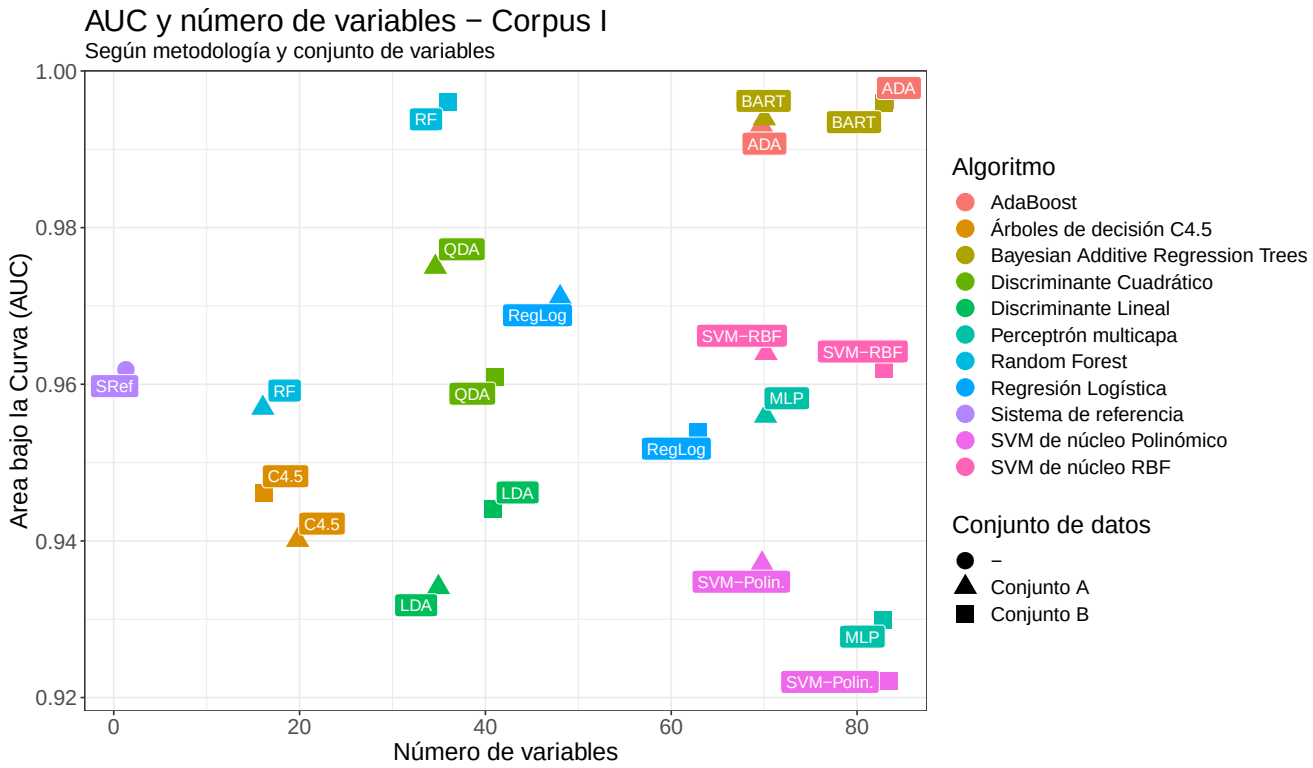

Figura 4.4: Área bajo la curva ROC y número de variables según el conjunto de datos del Corpus I usado y el clasificador empleado.

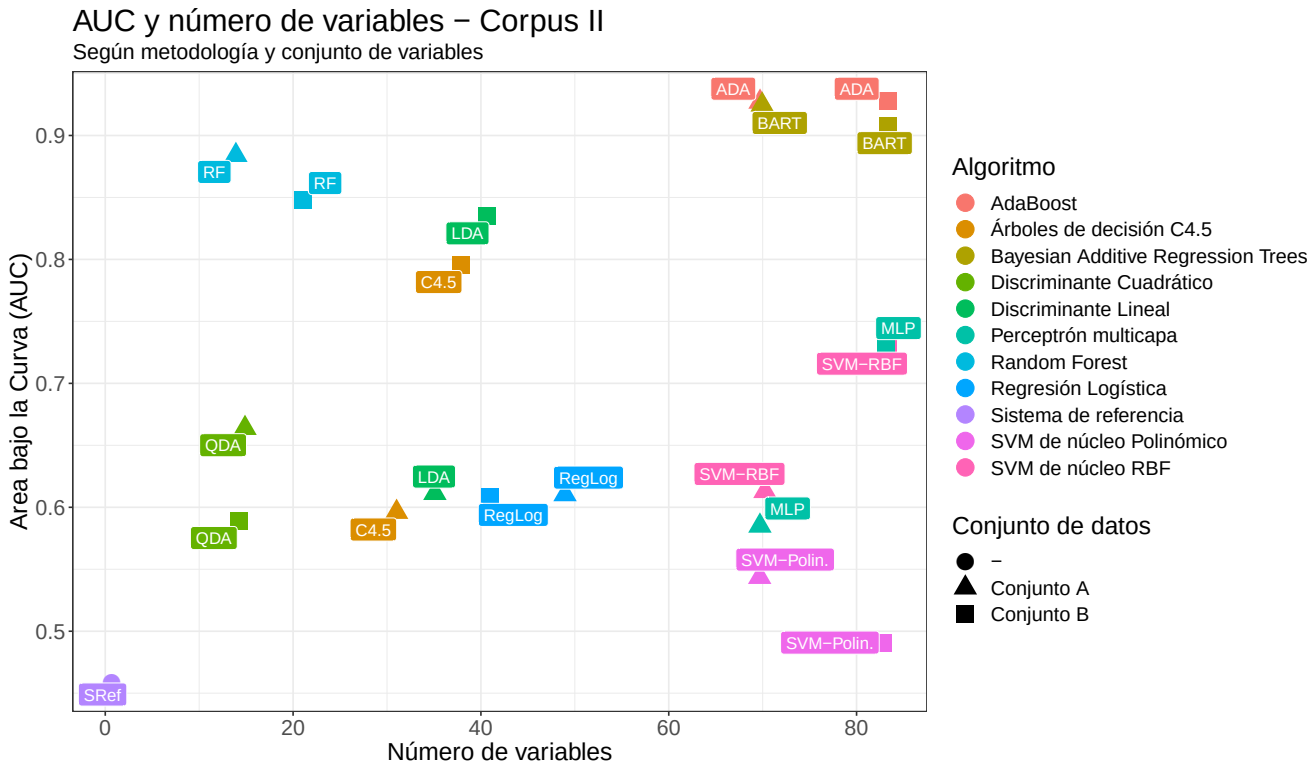

Figura 4.5: Área bajo la curva ROC y número de variables según el conjunto de datos del Corpus II usado y el clasificador empleado.

4.2. Comparativa de los modelos según precisión, especificidad y sensibilidad

La Tabla 4.2 muestra los resultados referentes a tasa de acierto, sensibilidad (True Positive Rate -TPR-) y especificidad (True Negative Rate -TNR-) de los modelos con el umbral probabilístico de clasificación asignado según el criterio desarrollado en el apartado correspondiente de los resultados de esta memoria. Como se ha comentado previamente, las estimaciones de estas medidas en los modelos de análisis supervisado se han realizado mediante validación cruzada de diez particiones.

Compendio general de resultados										
			Corpus I				Corpus II			
			Variables	Precisión	TPR	TNR	Variables	Precisión	TPR	TNR
	00	Sist. Ref.	1	95.75 %	95.80 %	95.34 %	1	45.03 %	49.50 %	44.90 %
04 An.Discr.	LDA	Conjunto A	35	95.29 %	98.28 %	68.12 %	35	97.15 %	4.67 %	99.69 %
		Conjunto B	41	94.08 %	98.29 %	55.71 %	41	96.67 %	28.03 %	98.56 %
	QDA	Conjunto A	35	95.93 %	96.09 %	94.50 %	15	96.85 %	1.87 %	99.46 %
		Conjunto B	41	95.53 %	95.79 %	93.23 %	14	97.27 %	1.87 %	99.90 %
05	C4.5	Conjunto A	20	97.89 %	98.88 %	88.85 %	31	95.24 %	28.75 %	97.97 %
		Conjunto B	16	98.09 %	98.99 %	89.84 %	38	96.69 %	58.75 %	98.25 %
06	RF	Conjunto A	16	98.45 %	99.18 %	91.82 %	14	98.42 %	60.00 %	100.00 %
		Conjunto B	36	**98.80** %	99.33 %	93.94 %	21	97.83 %	34.58 %	99.56 %
07	ADA	Conjunto A	70	98.29 %	91.96 %	98.99 %	70	**98.80** %	80.84 %	99.79 %
		Conjunto B	83	98.70 %	99.27 %	93.51 %	83	98.59 %	69.37 %	99.79 %
08	BART	Conjunto A	70	98.17 %	98.99 %	90.69 %	70	93.66 %	28.21 %	99.02 %
		Conjunto B	83	97.60 %	99.31 %	81.95 %	83	97.21 %	41.87 %	99.49 %

Tabla 4.2: Número de variables así como tasa de acierto, sensibilidad y especificidad estimada de los modelos estudiados en la memoria de estadística.

Dentro del Corpus I los resultados son excelentes ya que la precisión se mantiene siempre superior al 95 %. La sensibilidad, superior al 90 % en todos los modelos, se mantiene siempre por encima de la estimación de la especificidad. El sesgo de los clasificadores asignando la etiqueta 'defectuoso' en la predicción se debe principalmente a la existencia de más muestras con defectos que sin defectos dentro del Corpus I. Este hecho como es lógico afecta al entrenamiento y ajuste de los modelos.

El resultado base del sistema es bastante satisfactorio, logrando más del 95 % de tasa de acierto. El mejor resultado en lo que se refiere a tasa de acierto lo logra el modelo Random Forest con el juego de variables de los estadísticos. El modelo Adaptive Boost con este conjunto de variables logra una precisión menor por un margen pequeño. El conjunto de datos A con Random Forest, Adaboost y BART alcanza tasas de error inferiores al 2 % mientras que los árboles básicos C4.5 logran tasas de error cercanas al 3 %. LDA logra los modelos con peores resultados asociados de los modelos del Corpus I.

En el caso del Corpus II, la tasa de acierto no es una medida correcta para valorar la calidad discriminatoria del modelo. El desequilibrio en la distribución de clases de este Corpus (entre dos y tres imágenes con defectos por cada cien) da lugar a una tasa de acierto muy elevada con tan solo etiquetar todas las muestras como 'imagen sin defectos'. También hace que los clasificadores tengan un sesgo hacia la clase mayoritaria. Por lo tanto, será especialmente significativo valorar la sensibilidad o TPR en estos modelos.

El número de modelos que superan el umbral del 50 % de muestras con defectos bien clasificados es bajo; solo lo superan Random Forest con el conjunto de datos A, el árbol de decisión C4.5 inducido con el conjunto de datos B y los dos modelos de AdaBoost. El sistema de referencia queda próximo al 50 % pero éste, con su relativa sencillez, no logra una tasa de acierto satisfactoria. El modelo de Random Forest mencionado destaca por su relativa baja dimensionalidad (14 variables) y su aparente perfecta capacidad para distinguir tableros sin defectos. Los dos modelos AdaBoost logran los resultados más satisfactorios para el Corpus II con una elevada TPR que, en el caso del inducido a partir del conjunto de datos A, supera el 80 % de las clases positivas bien clasificadas. El árbol de decisión C4.5 con el conjunto B logra también una tasa de error baja junto a una sensibilidad cercana al 60 %. BART, el perceptrón multicapa y Random Forest, los tres con el conjunto B, parecen ser los siguientes modelos más eficaces clasificando aunque no superan el 50 % de los tableros con defectos bien clasificados. El resto de modelos al contar con sensibilidades tan bajas, no parecen tener utilidad práctica en el problema.

En las Figuras 4.6 y 4.7 se han representado mediante gráficos de líneas las medidas de capacidad de clasificación (precisión, sensibilidad y especificidad) de los modelos de análisis de datos expuestos relativos a los conjuntos de datos A y B compuestos por las características extraídas del Corpus I y II. A pesar de la similitud existente entre las Figuras, cabe destacar la notoria diferencia en rango de valores.

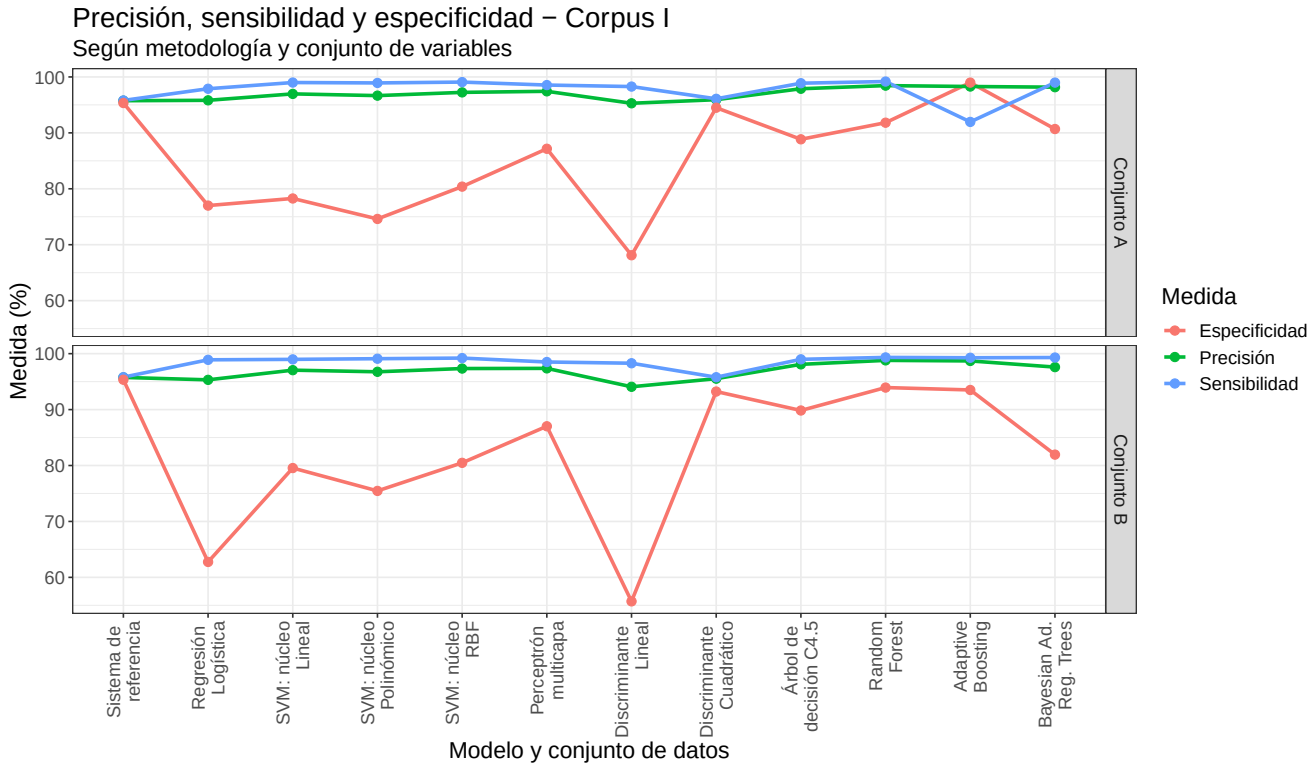

Figura 4.6: Precisión, sensibilidad y especificidad para cada uno de los conjuntos de datos del Corpus I y metodología empleada.

En el caso del Corpus I se observa cómo de forma general los resultados son muy altos exceptuando la especificidad de ciertos modelos (LDA). Los modelos en los que las tres tasas representadas han superado el 90 % han sido el sistema de referencia, los modelos de análisis discriminante cuadráticos y la mayoría de los modelos de agregación de árboles de decisión.

La Figura correspondiente al Corpus II evidencia la baja sensibilidad de los modelos. Existe una clara pobreza en los modelos para la clasificación de aquellas imágenes de tableros defectuosos. Tan solo los cuatro modelos que destacamos previamente superan el 50 % de instancias defectuosas bien clasificadas.

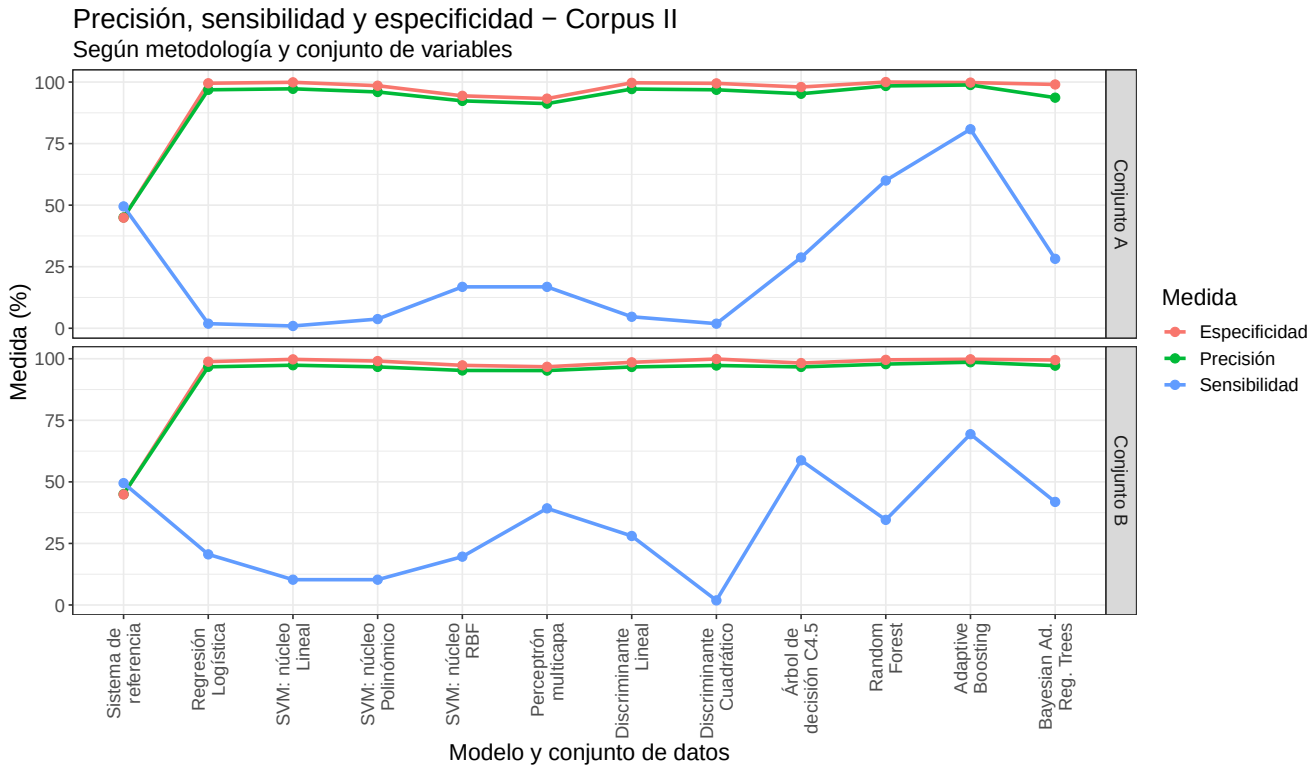

Figura 4.7: Precisión, sensibilidad y especificidad para cada uno de los conjuntos de datos del Corpus II y metodología empleada.

4.3. Discusión sobre modelos

En este apartado se han realizado representaciones en forma de malla o Spiderplots del ranking alcanzado por cada uno de los modelos con respecto al total de los mismos creados con los datos de ese Corpus en ambas memorias. A mayor área ocupada por los modelos representados, mayor calidad tendrá el clasificador en cuestión con respecto al resto. Si se diese un empate entre dos modelos en una medida, el ranking asignado será el que corresponde a la posición máxima. En la representación se han escogido cinco medidas tomadas en los modelos:

- Número de variables. Se trata de un ranking decreciente: a mayor número de atributos en el modelo, menor ranking asignado al modelo. Esta medida es de menor importancia con respecto al resto, principalmente por dos razones:

  1. Existen numerosos modelos en los que la medición requiere de que se extraigan todas las variables (por ejemplo, al necesitar de una componente alta extraída de la DCT).
  2. La clasificación de muestras no observadas en el conjunto de entrenamiento son prácticamente instantáneas en todos los modelos estudiados.

- Precisión. Se trata de un ranking creciente: los modelos cuya precisión sea más elevada, obtendrán un mayor ranking.

- Área bajo la curva ROC. Se trata de un ranking creciente: con más AUC, mayor ranking.

- F-Score. Tal y como se detallo previamente en la metodología, es trata de la media armónica de la sensibilidad y especificidad. Se trata de un ranking creciente: los modelos con mayor F-Score obtendrán un mayor ranking.

- Tasa de equierror. Se trata de un ranking decreciente: cuanto más EER tenga el modelo, menor resultará ser el ranking del modelo.

4.3.1. Corpus I

En la Tabla 4.3 aparecen reflejados los rankings, de mejor a peor, en cada uno de las medidas de los modelos previamente desarrollados para el Corpus I ordenados por la puntuación del F-Score. Esta medida se ha elegido para la ordenación al carecer completamente de valores faltantes y empates.

Ranking general de los modelos estudiados (Corpus I)						
	Conjunto	F-Score	Precisión	AUC	EER	Nº Variables
Random Forest	Conjunto B	23	23	23	23	17
Adaptive Boosting	Conjunto B	22	22	23	23	5
Sistema de referencia	-	21	5	15	18	23
Random Forest	Conjunto A	20	21	12	16	22
Adaptive Boosting	Conjunto A	19	20	19	20	10
Discriminante cuadrático	Conjunto A	18	7	18	18	19
Bayesian Additive Reg. Trees	Conjunto A	17	19	20	20	10
Discriminante cuadrático	Conjunto B	16	4	13	12	14
Árboles de decisión C4.5	Conjunto B	15	18	9	13	22
Árboles de decisión C4.5	Conjunto A	14	17	7	9	20
Bayesian Additive Reg. Trees	Conjunto B	11	16	23	21	5
Discriminante lineal	Conjunto A	3	2	5	5	19
Discriminante lineal	Conjunto A	1	1	8	7	14

Tabla 4.3: Ranking de los criterios seleccionados (Número de variables, Precisión, AUC, F-Score y EER) para los modelos estudiados en el Corpus I desarrollados en la memoria de estadística.

En las Figuras 4.8 y 4.9 aparece las representaciones de malla de los ranking del sistema de referencia y los modelos desarrollados en la memoria de estadística.

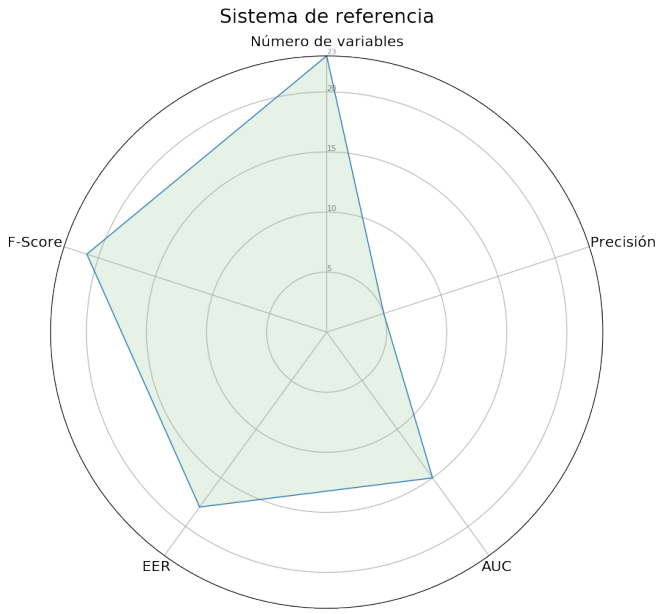

Figura 4.8: Representación gráfica del ranking en distintos criterios (Número de variables, Precisión, F-Score, EER y AUC) del sistema de referencia del Corpus I.

Figura 4.9: Representación gráfica del ranking en distintos criterios (Número de variables, Precisión, F-Score, EER y AUC) del sistema de referencia y los modelos desarrollados en la memoria de estadística a partir de los conjuntos de datos A y B del Corpus I.

Podemos hacer los siguientes comentarios en base a la representación de malla de los modelos del Corpus I:

- **Sistema de referencia**: en el Corpus I el sistema de referencia logra resultados bastante buenos: emplea el menor número posible de variables (una) y con ello consigue puntuaciones altas en todas las medidas a excepción de la precisión general del modelo.
- **Análisis discriminante**: el discriminante lineal en el Corpus I destaca únicamente por su simpleza en cuanto al número de variables. Bajo el resto de criterios son modelos muy pobres. La flexibilidad adicional que permite la heterocedasticidad mejora los resultados del discriminante. El discriminante cuadrático con el conjunto A es claramente superior al del B logrando medidas por encima de la media en número de variables, F-Score, EER y AUC.
- **Árboles de decisión C4.5**: Los árboles de decisión logran con un subconjunto notablemente reducido de variables una precisión elevada y un F-Score medio. El conjunto de variables B da lugar a un clasificador claramente superior a aquel asociado al conjunto A.
- **Random Forest**: Los clasificadores Random Forest parecen abarcar más área que el resto de modelos. Bajo muchos de los criterios obtienen puntuaciones excelentes. El conjunto de variables B da lugar a un clasificador que logra el ranking más alto en cuanto a F-Score, Precisión, EER y AUC. El A logra resultados buenos en F-Score, precisión y EER con el segundo juego de variables más pequeño seleccionado -16 variables-.
- **Adaptive Boosting**: AdaBoost logra resultados también extraordinarios. El conjunto B con la totalidad de variables logra las mejores puntuaciones en EER y AUC (empatando con el modelo de Random Forest con el conjunto B) y las segundas mejores puntuaciones en F-Score y precisión. El conjunto A logra resultados más modestos con 16 variables menos.
- **Bayesian Additive Regression Trees**: Los modelos BART logran excelentes puntuaciones en EER y AUC. El modelo A es más equilibrado con respecto a todos los criterios mientras que el creado a partir del conjunto B obtiene un alto ranking en EER y AUC que no parece traducirse en resultados significativos clasificando como podemos ver con el ranking asignado bajo los criterios de F-Score y precisión.

4.3.2. Corpus II

La Tabla 4.4 refleja el ranking, de mejor a peor, de los modelos creados a partir de los datos del Corpus II ordenados según la puntuación del F-Score. La medida se ha escogida de nuevo debido a su carencia de valores faltantes y la ausencia de empates en la misma.

En las Figuras 4.10 y 4.11 aparece la representación de malla de los ranking del sistema de referencia y los modelos desarrollados en la memoria de estadística. La peor resolución de las imágenes del Corpus II genera ruido y variaciones en las características extraídas de las imágenes. Esto da lugar a una nueva disposición de rankings en los modelos empleados.

Podemos hacer los siguientes comentarios en base a la representación de malla de los modelos del Corpus II:

- **Sistema de referencia**: en el Corpus II logra con una única variable medidas de precisión, EER y AUC bajas. Sin embargo, su F-Score -relevante ante la baja especificidad de los modelos- es relativamente elevado.
- **Análisis discriminante**: en el caso del discriminante lineal se obtienen resultados medios. El conjunto A ofrece un clasificador de baja dimensionalidad con una precisión asociada elevada mientras que el B, con más variables, destaca por su medida en el F-Score, EER y AUC. Los modelos cuadráticos pueden lograr rankings elevados en número de variables y precisión, pero son de poca relevancia ante el bajo valor del F-Score. Esto es una señal de un pobre poder discriminatorio.
- **Árboles de decisión C4.5**: el modelo B logra resultados significativos con pocas variables en la tasa de equierror, área bajo a curva ROC y F-Score. El conjunto con las componentes principales extraídas parece ofrecer como única ventaja un menor número de variables.

Ranking general de los modelos estudiados (Corpus II)						
	Conjunto	F-Score	Precisión	AUC	EER	Nº Variables
Adaptive Boosting	Conjunto A	23	23	22	23	10
Adaptive Boosting	Conjunto B	22	22	23	22	5
Random Forest	Conjunto A	21	21	19	17	22
Árboles de decisión C4.5	Conjunto B	20	10	16	19	16
Bayesian Additive Reg. Trees	Conjunto B	19	16	20	20	5
Random Forest	Conjunto B	17	20	18	18	19
Sistema de referencia	-	16	1	3	3	23
Árboles de decisión C4.5	Conjunto A	15	6	8	6	18
Bayesian Additive Reg. Trees	Conjunto A	14	4	21	21	10
Discriminante lineal	Conjunto B	13	9	17	16	15
Discriminante lineal	Conjunto A	6	15	11	9	17
Discriminante cuadrático	Conjunto B	4	18	7	10	22
Discriminante cuadrático	Conjunto A	2	14	14	13	20

Tabla 4.4: Ranking de los criterios seleccionados (Número de variables, Precisión, AUC, F-Score y EER) para los modelos estudiados en el Corpus II desarrollados en la memoria de estadística.

- **Random Forest**: logra resultados altamente satisfactorios: el modelo con el conjunto A logra altos valores en todos los criterios con un subconjunto de variables bastante reducido. El modelo asociado al conjunto de los estadísticos destaca por su EER ligeramente, mientras que bajo el resto de medidas parece ser inferior.
- **Adaptive Boosting**: los modelos de AdaBoost se alzan como los poseedores de los mejores resultados en todas las medidas relativas a la clasificación: el conjunto de datos A logra el valor más alto de F-Score, precisión y EER mientras que el B maximiza el AUC. A cambio, la dimensionalidad de los modelos es elevada.
- **Bayesian Additive Regression Trees**: en el caso del Corpus II los BART no logran resultados acordes a su complejidad computacional. Son resultados medio-altos en todas las medidas en base al uso completo de las dimensiones de los conjuntos de datos.

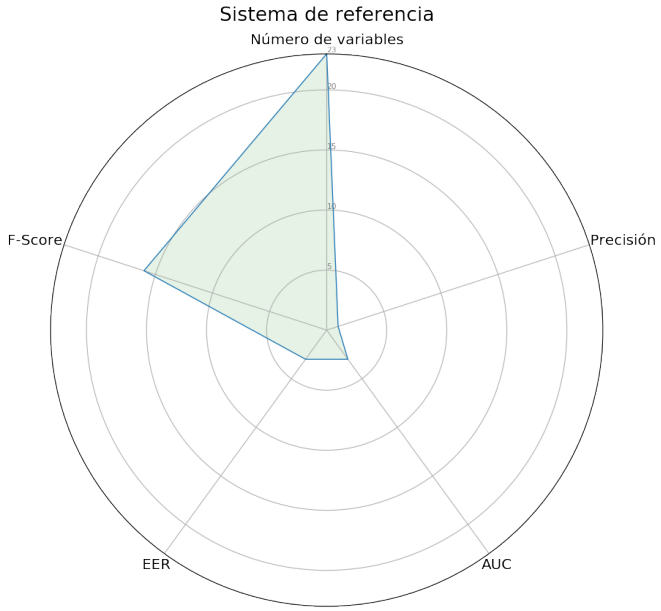

Figura 4.10: Representación gráfica del ranking en distintos criterios (Número de variables, Precisión, F-Score, EER y AUC) del sistema de referencia del Corpus II.

Figura 4.11: Representación gráfica del ranking en distintos criterios (Número de variables, Precisión, F-Score, EER y AUC) de los modelos desarrollados en la memoria de estadística a partir de los conjuntos de datos A y B del Corpus II.

4.4. Discusión de relevancia de las variables

En esta sección se realizará una breve discusión sobre la importancia relativa de las características que componen los conjuntos de variables que han servido de entrada a los modelos creados. La disquisición se realizará teniendo en cuenta las variables seleccionadas en la inducción de los modelos así como las figuras realizadas relativas a la importancia de variables en los modelos que lo permiten.

4.4.1. Variables relativas al filtrado

En la Tabla 4.5 podemos ver las variables que conforman este subconjunto de variables.

Variables relativas al filtrado	
AreaErrorB	AreaErrorN

Tabla 4.5: Variables directamente relacionadas con el filtrado.

La mayoría de los modelos inducidos con los datos del Corpus I emplean estas variables, especialmente la relacionada con el filtrado de negros. Variables relacionadas con la distribución de grises sirven de sustitutivos a estas variables. Esto se da de forma más frecuente en los modelos creados a partir del conjunto B de variables.

En el caso del Corpus II sucede lo contrario: son variables que se emplean mucho menos en modelos en los que se ha realizado una selección de variables o resultan ser variables a las que se les aporta una relevancia inferior. Los modelos de mejores resultados no destacan estas variables, por lo que parece que la menor resolución de las imágenes hace que este conjunto de características tenga más ruido y una menor componente de información útil.

4.4.2. Caracterizaciones de la distribución de grises

Componentes principales

En la Tabla 4.6 podemos ver los atributos que conforman este subconjunto de variables.

Distribución de grises: componentes principales				
- Img. original:	CompPrinOr1	CompPrinOr2	CompPrinOr3	CompPrinOr4
- Img. filtrada en blanco:	CompPrinTratB1	CompPrinTratB2		
- Img. filtrada en negro:	CompPrinTratN1	CompPrinTratN2		

Tabla 4.6: Atributos de componentes principales que caracterizan la distribución de grises.

En el caso del Corpus I resultan ser un conjuntos de características de alta utilidad para múltiples modelos. Se emplean principalmente las componentes extraídas de la imagen original, especialmente la 4 y 3. La segunda componente de la imagen procesada en negro se usa más que la primera en los modelos, aunque ninguna de las dos parece ser seleccionada en exceso o tener una importancia significativa. Muy posiblemente se deba esto a que el área de error relativa en negro aporta una información altamente similar a estas dos variables previamente mentadas. Las componentes extraídas de la imagen procesada en blanco se usan en menor medida en los modelos estudiados.

El Corpus II induce modelos en los que las componentes de la imagen original son muy relevantes o se seleccionan en multitud de modelos. La componente primera de la imagen tratada en negro resulta ser relevante y parece sustituir la información más pobre del área de error relativa del filtrado de negros. La segunda componente de la imagen procesada en blanco parece aportar información más útil que la primera. Muchos modelos requieren de información auxiliar a la aportada por estas variables, como la que pueden aportar las componentes de la DCT.

Estadísticos característicos

La Tabla 4.7 muestra aquellos atributos que forman parte de este conjunto de variables. El uso de los estadísticos, de forma general, parece requerir de un número de variables más alto que con el uso de las componentes principales.

Distribución de grises: estadísticos característicos				
-Img. original:	MediaOr	VarianzaOr	SkewnessOr	KurtosisOr
	MedianaOr	P10Or	P90Or	
-Img. filtrada en blanco:	MediaTratadaB	VarianzaTratadaB	SkewnessTratadaB	KurtosisTratadaB
	MedianaTratadaB	P10TratadaB	P90TratadaB	
-Img. filtrada en negro:	MediaTratadaN	VarianzaTratadaN	SkewnessTratadaN	KurtosisTratadaN
	MedianaTratadaN	P10TratadaN	P90TratadaN	

Tabla 4.7: Atributos estadísticos que caracterizan la distribución de grises.

Existe una variable predominante sobre el resto en todos los modelos inducidos, ya sean del Corpus I o del II, con este conjunto de datos: la kurtosis de la imagen tratada en negro. Es una característica que parece tener un papel esencial en los modelos cuyos resultados son los óptimos en cada uno de los Corpus.

Dentro de los modelos del Corpus I parece que las otras dos kurtosis, la de la imagen original y la de la imagen tratada en blanco, son empleadas frecuentemente. Esto también ocurre con las variables de apuntamiento o skewness. Se usan a menudo en los modelos también la media de la imagen original y otras como SkewnessTratadaN y VarianzaTratadaN. Dentro del subconjunto de variables extraídas de la imagen tratada en blanco, la más empleada es la mediana. Este subgrupo de variables es poco seleccionado o no se le otorga excesiva importancia dentro de los clasificadores. Un posible motivo puede ser la relativa escasez de muestras de defecto blanco.

En el Corpus II vuelven a destacar las variables relativas a apuntamiento y kurtosis de cualquiera de las imágenes con las que se ha trabajado. Se emplean en múltiples modelos entre los cuales se hallan los modelos que han logrado los mejores resultados. La media y mediana de la imagen original desatacan en muchos modelos así como ocurre con la varianza de la imagen tratada en blanco. Parece que existe diversidad de modelos en los que, aparte de las variables comentadas, se necesitan atributos que aporten información auxiliar (como las componentes de la DCT).

4.4.3. Componentes extraídas de la DCT

En la Tabla 4.8 refleja aquellas variables que componen el subgrupo de atributos.

Componentes extraídas de la DCT	
Img. filtrada sin binarizar:	Componentes de la 1 a la 30 extraídas.
Img. filtrada binarizada:	Componentes de la 1 a la 30 extraídas.

Tabla 4.8: Componentes de la DCT extraídas de la imagen tratada binarizada y sin binarizar.

Los modelos creados con los datos del Corpus I parecen reflejar que la información aportada por las componentes de la DCT es una adición a la que dan los estadísticos, las componentes principales y las áreas de error relativas. Las componentes tanto de la imagen binarizada como de la no-binarizada se seleccionan por igual en los modelos estudiados.

El Corpus II da lugar a modelos muy dependientes de la información aportada por las componentes DCT de la imagen tratada sin binarizar. La manifestación de este hecho es más agudo en el caso de inducirse los modelos con el conjunto de variables de los estadísticos. El caso más extremo es el modelo BART con este conjunto que, dando un peso muy fuerte a las componentes DCT de la imagen sin binarizar, logra uno de los mejores resultados. Las componentes de la imagen binarizada se usan de forman muy aislada y no se las da un peso de relevancia en ninguno de los clasificadores.

4.5. Conclusiones

El claro desbalance entre clases en los conjuntos de datos parece indicarnos que las medidas estimadas de los modelos más útiles para su comparación han sido la tasa de equierror, el área bajo la curva ROC y la sensibilidad y especificidad (o F-Score). La tasa de acierto resulta ser insuficiente, especialmente ante la aguda ausencia de imágenes con defectos en el Corpus II.

En cuanto a modelos, los mejores resultados provienen de clasificadores agregados cuyo clasificador base es un árbol de decisión. AdaBoost y Random Forest dan lugar a modelos con altas estimaciones en área bajo la curva ROC, tasa de equierror y medidas de acierto clasificando. Mientras que AdaBoost parece lograr resultados ligeramente superiores que Random Forest bajo múltiples criterios, éste alcanza resultados excelentes con un número de variables bajo. La existencia de más ruido y mayor desbalance de clases en el Corpus II pone de manifiesto que los modelos basados en la agregación de árboles de decisión son considerablemente robustos.

Con respecto a las características extraídas de las imágenes, podemos destacar como el área de error relativa en negro ha resultado ser en en general una variable útil en la mayoría de modelos, mientras que la de blancos no. Esto podría ser reflejo de que el filtrado de blancos no ha sido tan efectivo. En cuanto a las variables relacionadas con las distribuciones de grises, parece que todas las componentes principales mantienen una importancia relativa similar mientras que en los estadísticos característicos parece destacar la kurtosis de la imagen tratada en negro. Esta variable podría tener el mayor poder discriminatorio de todas las estudiadas. Las componentes de la DCT parecen aportar información útil y complementaria a la aportada por los otros dos grupos de variables. La combinación adecuada de estas puede dar lugar a clasificadores cuya discriminación es excelente como vemos en el caso del mejor modelo del Corpus II (AdaBoost con el conjunto A).

Capítulo 5

Conclusiones y trabajo futuro

En este capítulo se expondrán las conclusiones alcanzadas con la finalización del proyecto.

En una primera sección se relacionará el proyecto con los contenidos de las asignaturas del grado. Después se realizará un balance respecto a la consecución de los objetivos para dar paso finalmente a un breve desarrollo sobre posibles líneas de trabajo posteriores a la finalización del proyecto.

5.1. Relación con los contenidos del grado

Las asignaturas del grado correspondientes al análisis de datos ('Análisis de datos' y 'Análisis Multivariante') han dado el fundamento esencial necesario para el desarrollo del proyecto. Cabe destacar las nociones de regularización adquiridas en otras asignaturas ('Regresión y Anova' y 'Modelos Lineales').

Asignaturas como 'Inferencia estadística' y 'Computación Estadística' sirven como base en los intervalos de confianza calculados en los resultados y en el manejo general del lenguaje de programación R. La importancia dada al equilibrio de las muestras en cualquier estudio proviene de 'Muestreo Estadístico I' y la relevancia de las medidas propias de las clasificaciones binarias de 'Análisis de Datos Categóricos'. Procedimientos propios para el desarrollo de los modelos más avanzados como el muestreo de Gibbs se impartieron en las asignaturas 'Modelos Estadísticos Avanzados' y 'Métodos Estadísticos de Computación Intensiva'.

5.2. Objetivos alcanzados

El correcto desarrollo del proyecto nos ha permitido lograr una serie de objetivos:

- Los dos Corpus de imágenes, Corpus I (conjunto de imágenes con una mayoría de imágenes con defectos y de alta resolución) y Corpus II (conjunto de imágenes de menor resolución con pocas imágenes con defectos), se han clasificado mediante visualización en imágenes sin defectos o con defectos negros, blancos o topográficos. Ha resultado ser la parte más ardua del proyecto en cuanto a horas dedicadas de trabajo.

- De las imágenes filtradas con el procedimiento descrito en la memoria de ingeniería informática [1] se han extraído tres grupos de variables o características (áreas de error relativas, caracterizaciones de las distribuciones de grises y componentes de la DCT) que han servido de entrada para los modelos de análisis supervisado planteados posteriormente.

- Se han ajustado correctamente modelos de análisis discriminante, árboles de decisión y la agregación de estos últimos. Se ha trabajado con modelos de discriminante lineal con regularización L1 y con discriminante cuadrático. Se han inducido árboles de decisión estilo C4.5 ajustando el valor de los parámetros de coste e instancias por nodo terminal. En el caso de la agregación de árboles, se han valorado modelos de Bagging (Random Forest con selección de variables) y Boosting. Los modelos estudiados de esta última metodología han sido Adaptive Boosting y Bayesian Additive Regression Trees. Sobre estos modelos se ha estudiado la influencia de diversos parámetros como el número de árboles básicos o aquellos que controlan la regularización.

- La discusión de resultados nos ha permitido valorar la calidad de discriminación de los modelos desarrollados en esta memoria de tal forma que se pudiese valorar cual era el sistema óptimo según diversos criterios. Se ha valorado la relevancia de cada uno de los grupos de variables entrada a los modelos en cuanto a su aporte de información útil a los mismos. La disquisición se ha realizado acompañándola de gráficos que facilitasen la comparativa.

- Con el desarrollo del proyecto en su conjunto se han podido crear modelos de análisis supervisado de agregación de árboles de decisión capaces de discriminar con alta precisión entre imágenes con defectos de las que carecen de ellos. Los modelos de Random Forest y AdaBoost logran resultados altamente satisfactorios en las diferentes medidas relativas a su calidad como clasificadores. Aquellos creados a partir del conjunto de variables caracterizado por el uso de estadísticos combinan estos con la información extraída de la DCT. Destacan por encima del resto de variables la Kurtosis de la imagen tratada en negro, dato que concuerda con la relación existente entre esta medida y el número de puntos extraños en una distribución. Por otro lado, aquellos inducidos con el conjunto de datos de las componentes principales parecen marcar como relevantes estas componentes. Aquellas extraídas a partir de la transformada discreta del coseno (DCT) aportan información de utilidad pero complementaria en comparación con componentes principales en estos modelos.

5.3. Trabajo Futuro

En cuanto a posible trabajo futuro relativo a este Trabajo de Fin de Grado, planteamos tres posibles ramas de desarrollo:

- Un nuevo conjunto de datos facilitado por la empresa podría servirnos de conjunto de prueba sobre el que validar con mayor certeza el que nuestros sistemas son más precisos que el instalado en la factoría ('Smart Eyes'). Otra posibilidad sería el poner en funcionamiento los modelos en la propia línea de producción.

- Los resultados estimados de los modelos parecen evidenciar que con equipos de visión artificial más modernos se podrían alcanzar tasas de acierto cercanas al 100 %. Las imágenes empleadas en el desarrollo del TFG provienen de cámaras con más de 5 años de antigüedad que ya en el momento de su compra estaban hasta cierto punto obsoletas.

- Por otro lado se podría continuar trabajando con los conjuntos de imágenes ya facilitados por la empresa y aplicarles diversos modelos de agrupación de clasificadores más complejos. Otra opción sería aplicar modelos de análisis no supervisado o clustering que agrupasen automáticamente los datos en diversos grupos que se pudiesen identificar como imágenes con defectos o sin ellos.

Bibliografía

[1] Alejandro Rodríguez Collado, Mª Aránzazu Simón Hurtado y Carlos Enrique Vivaracho Pascual. *Detección de defectos en tiempo real en una línea de fabricación de tableros mediante técnicas de reconocimiento de patrones*. Departamento de Informática, Escuela de Ingeniería Informática. Universidad de Valladolid, 2019.

[2] T.A. Banet, T. Aluja y A. Morineau. *Aprender de los datos: el análisis de componentes principales : una aproximación desde el Data Mining*. Ciencia y tecnología. EUB, 1999. ISBN: 9788483120224. URL: https://books.google.es/books?id=6p7mPQAACAAJ.

[3] Gilbert Strang. "The Discrete Cosine Transform". En: *SIAM Rev.* 41.1 (mar. de 1999), págs. 135-147. ISSN: 0036-1445. DOI: 10.1137/S0036144598336745. URL: http://dx.doi.org/10.1137/S0036144598336745.

[4] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN: 9780471458760. URL: https://books.google.es/books?id=hpEzw4T0sPUC.

[5] Luis Ángel García y Miguel Alejandro Fernández. *Apuntes de la asignatura "Análisis de Datos"*. Departamento de Estadística e Investigación Operativa, Facultad de Ciencias. Universidad de Valladolid, 2016.

[6] Teodoro Calonge Cano y Carlos J. Alonso González. *Apuntes de la asignatura "Técnicas de Aprendizaje Automático"*. Departamento de Informática, Escuela de Ingeniería Informática. Universidad de Valladolid, 2018.

[7] Teodoro Calonge Cano y Carlos J. Alonso González. *Apuntes de la asignatura "Minería de Datos"*. Departamento de Informática, Escuela de Ingeniería Informática. Universidad de Valladolid, 2018.

[8] Viv Bewick, Liz Cheek y Jonathan Ball. "Statistics review 13: receiver operating characteristic curves". En: *Critical care (London, England)* 8.6 (2004), 508—512. ISSN: 1364-8535. DOI: 10.1186/cc3000. URL: http://europepmc.org/articles/PMC1065080.

[9] Suzanne Ekelund AcuteCareTesting. *ROC curves – what are they and how are they used?* 2011. Último acceso: 23-11-2018. URL: https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used.

[10] Daniel Peña. *Análisis de datos multivariante*. Madrid: McGraw-Hill, 2002. ISBN: 9788448136109. URL: https://books.google.es/books?id=TrVlAAAACAAJ.

[11] William D. Neal. *Using Discriminant Analysis in Marketing Research: Part 1*. 1989. Último acceso: 23-11-2018. URL: https://archive.ama.org/archive/ResourceLibrary/MarketingResearch/documents/6896090.pdf.

[12] Suzanne Ekelund AcuteCareTesting. *Linear discriminant analysis reveals differences in root architecture in wheat seedlings related to nitrogen uptake efficiency*. 2017. Último acceso: 23-11-2018. URL: https://academic.oup.com/jxb/article/68/17/4969/4243598.

[13] Kunlun Li y col. "Multi-class text categorization based on LDA and SVM". En: *Procedia Engineering* 15 (2011). CEIS 2011, págs. 1963 -1967. ISSN: 1877-7058. DOI: https://doi.org/10.1016/j.proeng.2011.08.366. URL: http://www.sciencedirect.com/science/article/pii/S1877705811018674.

[14] Scikits-Learns. *Linear and Quadratic Discriminant Analysis with confidence ellipsoid*. 2016. Último acceso: 05-08-2018. URL: https://preview.tinyurl.com/yaxasgud.

[15] GitHub. *R caret: rfe nnet "undefined columns selected*. 2016. Último acceso: 21-09-2018. URL: https://github.com/topepo/caret/issues/485.

[16] David Cournapeau. *Scikit-learn*. 2018. Último acceso: 20-06-2018. URL: http://scikit-learn.org/stable/.

[17] David Cournapeau. *Scikit: RFE*. 2018. Último acceso: 30-07-2018. URL: https://medium.com/@aneesha/recursive-feature-elimination-with-scikit-learn-3a2cbdf23fb7.

[18] Stack Exchange. *How to interpret a VIF higher than 4?* 2017. Último acceso: 04-08-2018. URL: https://stats.stackexchange.com/questions/169445/how-to-interpret-a-vif-of-4.

[19] Alan Julian Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 1.ª ed. Springer Publishing Company, Incorporated, 2008. ISBN: 0387781889, 9780387781884.

[20] Luis Ángel García y Miguel Alejandro Fernández. *Apuntes de la asignatura "Análisis de Datos Multivariantes"*. Departamento de Estadística e Investigación Operativa, Facultad de Ciencias. Universidad de Valladolid, 2017.

[21] Sherman Horn. *Decision trees: from ID3 to C4.5*. 2012. Último acceso: 14-10-2018. URL: https://slideplayer.com/slide/7606786/.

[22] Max Kuhn. *Package 'caret' Manual*. New London: CRAN, 2018.

[23] Kurt Hornik. *Package 'RWeka' Manual*. Viena: CRAN, 2018.

[24] John Ellson. *Graphviz - Graph Visualization Software*. 2018. Último acceso: 12-10-2018. URL: https://www.graphviz.org/.

[25] R.; Friedman J. Hastie T.; Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

[26] Global Software Report. *Random Forest Classifier – Machine Learning*. 2017. Último acceso: 05-08-2018. URL: http://www.globalsoftwaresupport.com/random-forest-classifier-bagging-machine-learning/.

[27] Katherine Gray y col. "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease". En: *NeuroImage* 65 (oct. de 2012). DOI: 10.1016/j.neuroimage.2012.09.065.

[28] Fortran original by Leo Breiman, R port by Andy Liaw Adele Cutler y Matthew Wiener. *Package 'randomForest' Manual*. CRAN, 2018.

[29] Brian Lehman. *AdaBoost – Machine Learning in Action*. 2015. Último acceso: 14-10-2018. URL: https://github.com/DrSkippy/Data-Science-45min-Intros/tree/master/adaboost-101.

[30] Reinhard Klette. *Concise Computer Vision - An Introduction into Theory and Algorithms*. Undergraduate Topics in Computer Science. Springer, 2014, págs. 1-413. ISBN: 978-1-4471-6319-0. URL: http://dx.doi.org/10.1007/978-1-4471-6320-6.

[31] Matias Gamez-Martinez Esteban Alfaro-Cortes y Noelia Garcia-Rubio. *Package 'adabag' Manual*. Albacete: CRAN, 2018.

[32] Justin Bleich Adam Kapelner. *Package 'bartMachine' Manual*. New York: CRAN, 2017.

[33] Zak Jost. *Bayesian Additive Regression Trees Paper Summary*. 2017. Ultimo acceso: 09-04-2018. URL: https://towardsdatascience.com/bayesian-additive-regression-trees-paper-summary-9da19708fa71.

[34] B. Hernández y col. "Bayesian Additive Regression Trees using Bayesian model averaging". En: *Statistics and Computing* 28.4 (2018), págs. 869-890. ISSN: 1573-1375. DOI: 10.1007/s11222-017-9767-1. URL: https://doi.org/10.1007/s11222-017-9767-1.

[35] Junni L. Zhang y Wolfgang K. Härdle. "The Bayesian Additive Classification Tree applied to credit risk modelling". En: *Computational Statistics and Data Analysis* 54.5 (2010), págs. 1197 -1205. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2009.11.022. URL: http://www.sciencedirect.com/science/article/pii/S0167947309004356.