EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

RESUMEN

La ciencia ha desarrollado múltiples aplicaciones que buscan estudiar las coberturas del suelo apoyadas en la clasificación de productos provenientes de sensores remotos. Para ello, se aplican métodos de clasificación de imágenes satelitales, como: Arboles de Decisión (AD), Maquinas de Soporte Vectorial (SVM), Random Forest y Mahalanobis, clasificaciones objeto de estudio de este trabajo. Para el estudio se utilizo una imagen a color Ultracam de la ciudad de Bogotá de 2007, compuesta de tres bandas. En el desarrollo del trabajo se logro obtener porcentajes correctamente clasificados por encima del 85%, que garantiza la obtención de clasificaciones apropiadas.

Palabras Clave: Reconocimiento de patrones, Arboles de Decisión, SVM, Mahalanobis, Random Forest.

1. INTRODUCCION

El origen de la percepción remota se define textualmente como la tecnología que permite la adquisición de información de objetos, sin tener un contacto físico con ellos, para el estudio de la cobertura vegetal y el uso de la tierra, los sensores remotos juegan un papel muy importante en términos de la adquisición de datos, por la capacidad que ofrecen para obtener información multitemporal determinada por la frecuencia de la toma de datos, la cual posibilita la clasificación [1]. Se han desarrollado múltiples aplicaciones que buscan estudiar las coberturas del suelo apoyadas en la clasificación de productos provenientes de sensores remotos que regularmente utilizan la agrupación de píxeles, a partir de parámetros estadísticos y códigos matemáticos que arreglados en un programa permiten analizar utilizando los métodos: Mahalanobis, los arboles de decisión, Maquinas de Soporte Vectorial (SVM) y Random Forest por su traducción del inglés Bosques Aleatorios (BA) Cada uno contiene:

la distancia Mahalanobis, que utiliza valores espectrales medios de cada clase y sus covarianzas.

Por su parte, SVM, aplica funciones Kernel que definen hiperplanos de separación entre las clases [2]. Por ejemplo, la Función Base Radial Gaussiana contiene los parámetros: C, que es una cantidad que penaliza los vectores de soporte clasificados en otra clase, y sigma es un coeficiente de potencia que varía de acuerdo a las especificaciones que se definan [3].

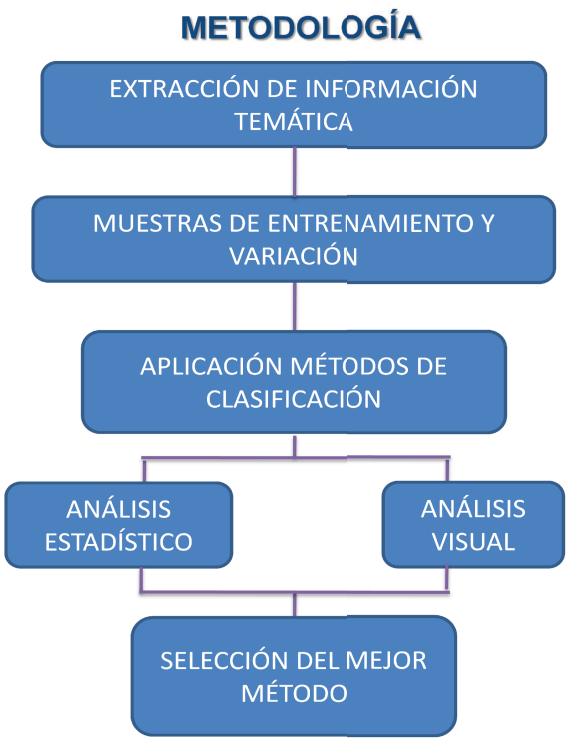
Por su parte, Los *arboles de decisión*, tiene criterios definidos por el intérprete implementando una estructura ramificada de “árbol” [4]. Para cada nivel del “arból” se definen intervalos de niveles digitales de acuerdo a la categoría que se busca diferenciar [5].

Random Forest (Bosques Alearorios, BA) es un método de clasificación y regresión basado en grandes números de árboles de decisión.

En el presente trabajo se busca comparar los métodos de clasificación: Arboles de Decisión (AD), Maquinas de Soporte Vectorial (SVM), Mahalanobis y Arboles Aleatorios (BA), y escoger la mejor clasificación.

2. METODOLOGIA

Imagen 1. Metodología



EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

2.1. EXTRACCION DE INFORMACION TEMATICA

2.1.1. Búsqueda de la información e insumos para el desarrollo del trabajo.

Se obtuvo una Imagen Digital Ultracam de la ciudad de Bogotá, por sus características se tienen tres bandas, para el manejo digital hubo que realizar un recorte quedando esta de un tamaño 3.35 millones (pixeles). La zona escogida para el estudio es en el sector centro – oriental de Bogotá cerca a los cerros orientales.

Imagen1. Zona de estudio

2.1.2. Ubicación de las Muestras de Entrenamiento.

La zona cubierta por la Imagen es una zona urbana ubicada entre la Carrera 10 y los Cerros Orientales, y por la Av. Calle 19 a la calle 26, las cuales se escogieron seis clases de cobertura:

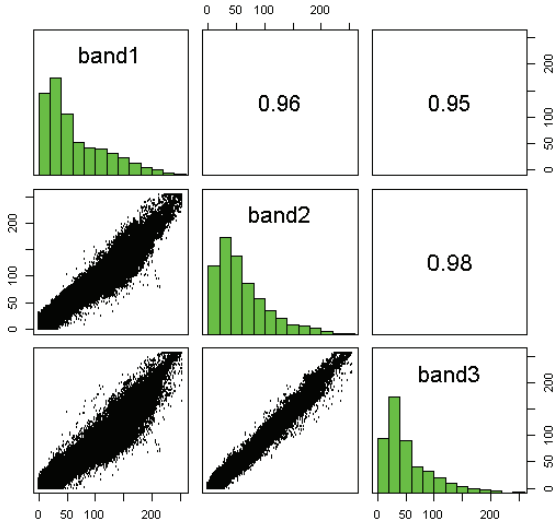
1. Bosques (entre la Cra. 7 y Cerros orientales),
2. Techos (cubiertas de edificaciones),
3. Sombras (proyecciones sobre el suelo), 4 Pastos (sobre canchas de futbol) y 5. Suelo desnudo (entre la Av. Circunvalar y los Cerros Orientales) y 6. Vías. Imagen 2.

Imagen2. Polígonos de entrenamiento

2.1.3. Muestras de Entrenamiento.

A través de ArcGis se digitalizarón los polígonos de entrenamiento y validación obteniendo el respectivo shapefile (“MUESTRAS”), se cargo la imagen en “R” [6], visualizando sus bandas, composición de color, matrices de covarianza y correlación como se presenta en las tablas 1, 2, 3. En donde la correlación es alta entre las diferentes bandas. Imagen 3.

Imagen3. Correlación entre las bandas



EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

Tabla 1.Resumen estadísticas por Banda

	band1	band2	band3
Min.	0	0	0
1st Qu.	24	26	24
Median	45	48	39
3rd Qu.	96	83	74
Max.	255	255	255
NA's	0	0	0

Tabla 2. Matriz de Covarianzas

	band1	band2	band3
band1	2816.109	2343.093	2287.450
band2	2343.093	2104.822	2049.226
band3	2 287.450	2049.226	2073.194

Tabla 3. Matriz de Correlación

	band1	band2	band3
band1	1.0000000	0.9624033	0.9466883
band2	0.9624033	1.0000000	0.9809849
band3	0.9466883	0.9809849	1.0000000

2.1.4. Adición de las Clases de Cobertura a los Puntos de Entrenamiento

El shapefile exportado de Arcgis (“MUESTRAS”), se carga en R, con el objeto de crear las correspondientes clases (spatial polygons dataframe), para cada pixel de la imagen, la clase de cobertura existente. Ver Imagen 4 y 5.

Imagen4. Obtención de clases existentes en los sitios de muestreo

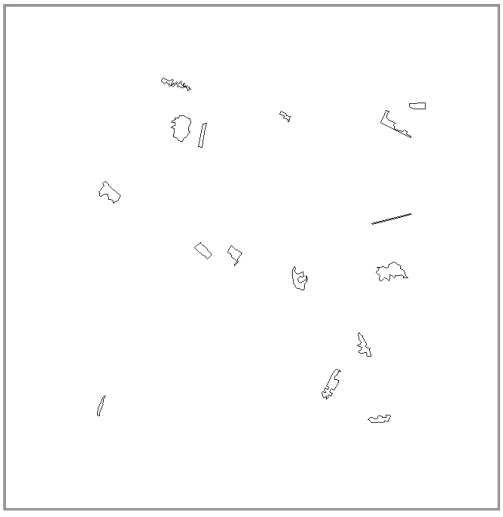
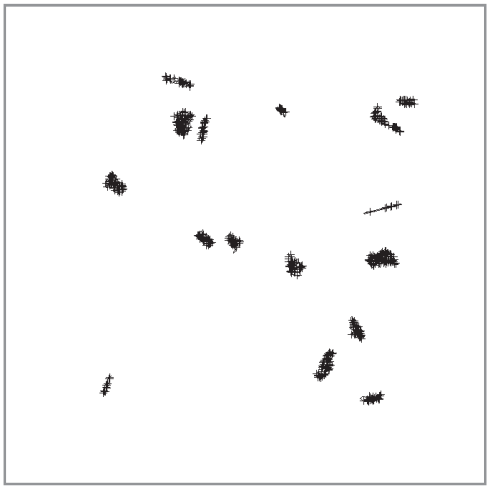


Imagen5. Adición de clases de cobertura



3. RESULTADOS

Los resultados de exactitud temática de está las diferentes se resumen en sus respectivas tablas. Éstas contienen: la matriz de confusión, exactitud temática PCC%, el Indice Kappa y los correspondientes intervalos de confianza.

3.1. CLASIFICACIÓN DE LA COBERTURA DEL SUELO USANDO DISTANCIA MAHALANOBIS.

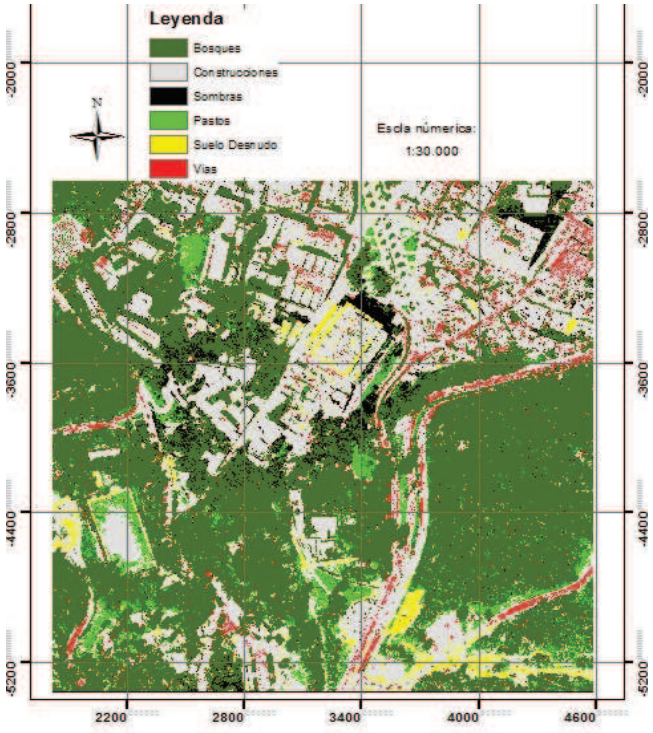
Tabla 4. Matriz de confusion 1

	true					
predicted	**1**	**2**	**3**	**4**	**5**	**6**
1	220	21	3	5	0	0
2	0	264	0	0	0	0
3	0	4	219	0	0	0
4	4	0	0	271	0	0
5	0	5	0	0	138	0
6	0	3	0	0	0	43

PCC 96.25
Kappa 0.953396
Intervalos de Confianza
Lim.sup
0.9718573
Lim_inf
0.9501909

EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

Imagen 6. Imagen Clasificada mediante DM



3.2 CLASIFICACIÓN DE LA COBERTURA USANDO ARBOLES DE DECISIÓN (AD)

Aquí se puede observar que la banda 5 fue la seleccionada como nodo principal. Ver imagen 6.

Imagen 6. Diagrama Árboles de Decisión.

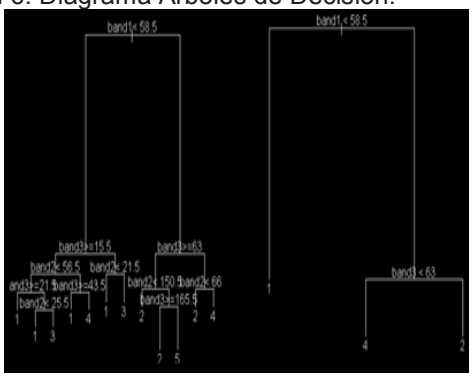
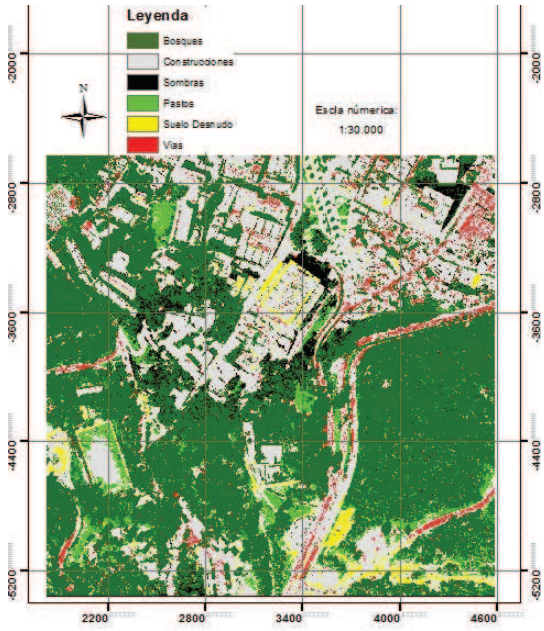


Tabla 5. Matriz de confusion 2

	true					
predicted	**1**	**2**	**3**	**4**	**5**	**6**
1	233	0	6	10	0	0
2	0	252	0	0	8	4
3	12	1	210	0	0	0
4	22	1	0	238	0	14
5	0	6	0	0	137	0
6	0	1	0	2	0	43

Pcc 92.75
Kappa 0.9103211
Intervalos de Confianza
Lim_sup
0.9408478
Lim_inf
0.9114238

Imagen 7. Imagen Clasificada mediante Arboles de decisión



3.3 CLASIFICACIÓN POR MEDIO DE SVM

Tabla 6. Matriz de confusion 3

	true					
predicted	1	2	3	4	5	6
1	240	0	3	6	0	0
2	0	262	0	0	1	1
3	3	1	219	0	0	0
4	6	0	0	269	0	0
5	0	1	0	0	142	0
6	0	0	0	0	0	46

PCC 98.16667
Kappa 0.9772524
Intervalos de Confianza
Lim_sup
0.9878623
Lim_inf
0.972397

EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

Imagen 8. Imagen Clasificada mediante SVM

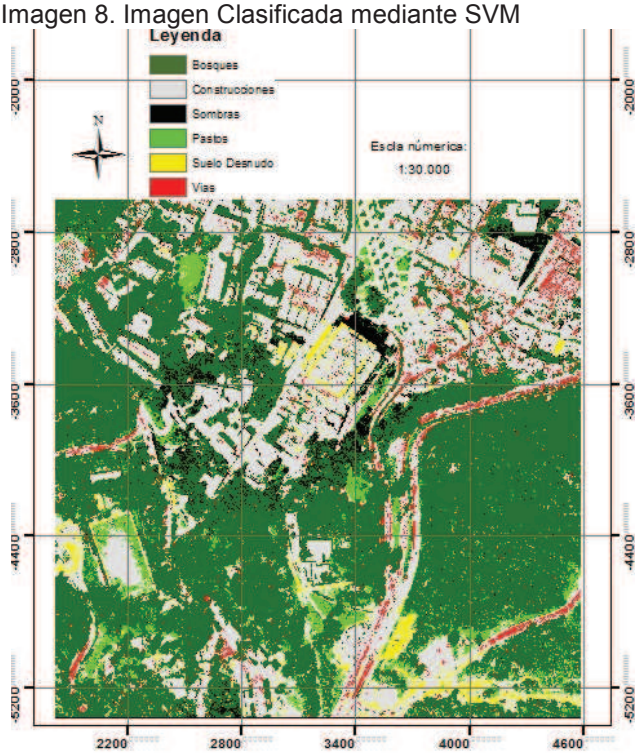
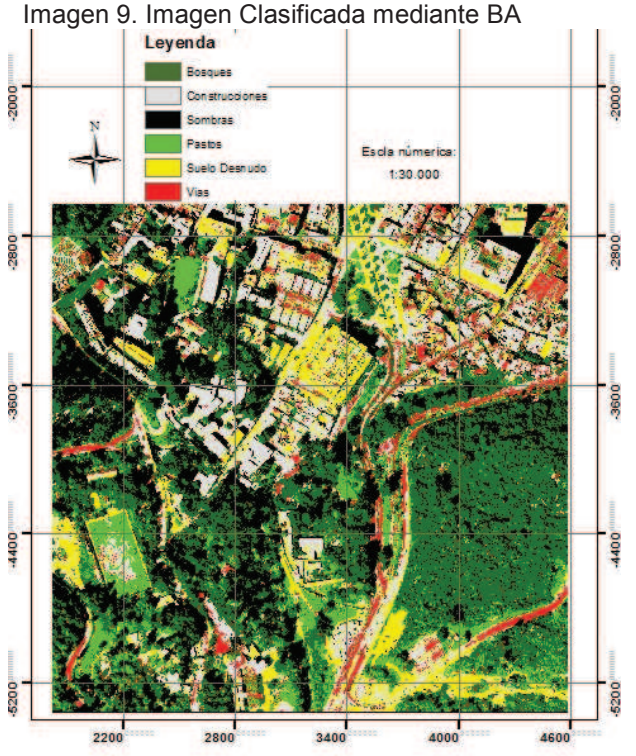


Imagen 9. Imagen Clasificada mediante BA



3.4 CLASIFICACIÓN POR MEDIO DE BOSQUES ALEATORIOS (BA)

Tabla 7. Matriz de confusion 4

	True					
Predicted	**1**	**2**	**3**	**4**	**5**	**6**
1	238	0	3	8	0	0
2	1	256	0	0	6	1
3	9	1	213	0	0	0
4	9	0	0	266	0	0
5	1	3	0	0	139	0
6	0	1	0	0	0	45

PCC 96.41667
Kappa 0.9555358
Intervalos de Confianza
Lim_sup
0.9732895
Lim_inf
0.9520815

Tabla 8. Resumen resultados

Método	**PCC%**	**índice Kappa**	**Lim. Sup.**	**Lim. Inf.**
Mahalanobis	96.25	*0.953*	*0.971*	*0.950*
AD	92.75	.910	0.940	0.911
SVM	98.16	0.977	0.987	0.972
BA	96.41	0.955	0.973	0.952

4. DISCUSION

4.1. Clasificación de la cobertura del suelo usando distancia mahalanobis.

Evaluación de exactitud temática: las clases obtenidas presentan un porcentaje correctamente clasificado (pcc) = 96.4%, con error de matriz de confusión de 3.6%. Frente a un indice Kappa (k1) = 95.5%, se observa igual nivel digital (ND) para las clases “techos” con las vías en algunos casos. Las demás clases presentan mayor diferencia entre ellas debido al dominio espectral que tienen es distinto.

4.2. ARBOLES DE DECISIÓN

A través de la predicción dada por la matriz de confusión (error =0.111), el pcc obtenido (88.8%) frente a la imagen clasificada (imagen 4), mediante los arboles de decisión se obtuvieron ganancias en la clasificación en un 3.6%, y en la precisión de imagen (kappa=82.7%) del 3.9% (ver imágenes 5 y 5a).

4.3. CLASIFICACIÓN POR MEDIO DE SVM

Al emplear las SVM, Se obtuvo similitud frente a la clasificación realizada con los arboles de decisión, se puede deducir que las SVM realizan análisis deductivos para los procesos de entrenamiento.

4.4. CLASIFICACIÓN POR MEDIO BOSQUES ALEATORIOS

Al emplear las Bosques Aleatorios, Se obtuvo una clasificación bastante discriminada y detallada. Las sombras del bosque fueron muy bien interpretas lo que da la sensación de una vista 3D. El resultado visualmente es muy diferente al presentado por los demás métodos.

Algunos autores plantean, en la etapa de extracción de información temática a partir de las imágenes, técnicas que conducen a la aplicación de metodologías inductivas o deductivas. La selección del tipo de metodología será función de si se parte de un diseño experimental para extraer leyes (metodología inductiva), o bien, si se proponen estimaciones al analizar las relaciones teóricas entre los componentes que intervienen en el problema (metodología deductiva). [7].

4. DISCUSION

Los métodos convencionales de clasificación, requieren de la experticia y conocimiento técnico en procesamiento digital de imágenes. La clasificación orientada objetos, necesita además, reconocer sus atributos geométricos y para lograrlo, se hace necesario contar con programas y metodologías que faciliten la aplicación de esta técnica.

Para obtener mayor precisión en la correcta clasificación (pcc), se debe tener cuidado en la determinación de las diferentes clases de cobertura, ya que al no tener una buena diferenciación en la delimitación de las clases, los resultados serán homogéneos, es decir no se pueden diferenciar en el mapa los tipos de cobertura.

Es bien importante la Clasificación de la cobertura usando los Arboles de decisión, ya que se obtienen ganancias en la clasificación y en la precisión de la imagen (mapa).

Las SVM aportan buenos resultados en la extracción de información temática, se puede observar una alta similitud con la Clasificación de la cobertura realizada por medio de los Arboles de decisión. (pcc=88.8, kappa=82.7).

Se puede concluir que una de las mejores clasificaciones para la cobertura del suelo, se obtiene mediante el uso de los árboles de decisión con sus respectivas podas, en el caso de estudio, se obtuvo un resultado similar para SVM.

El mejor algoritmo para discriminar información fue Bosques aleatorios, por lo cual, se selecciono como el mejor método de clasificación en este trabajo.

Los métodos Mahalanobis, Árboles de Decisión y SVM presentaron un buen nivel de homogenización de la información.

La selección de una adecuada muestra es importante. De la pureza de ésta depende el éxito de la clasificación.

Una buena cantidad de puntos de entrenamiento y de verificación sirvieron bastante en la clasificación obtenida, por lo que se recomienda mantener una relación de un tercio (1/3) para entrenamiento y (2/3) para la muestra sobre el total de la muestra.

EXTRACCION DE INFORMACION TEMATICA A PARTIR DE LOS MÉTODOS DE CLASIFICACIÓN: MAHALANOBIS, ARBOLES DE DESICIÓN, SVM Y BOSQUES ALEATORIOS SOBRE UNA IMAGEN ULTRACAM (2007), DEL SECTOR CENTRO - ORIENTE DE BOGOTA D.C.

REFERENCIAS

[1] Metodología para detectar cambios en el uso de la tierra utilizando los principios de la clasificación orientada a objetos, estudio de caso piedemonte de Villavicencio, Meta, Andrés Felipe Rodríguez Vásquez. Facultad de Ingeniería Agronómica, Universidad Nacional de Colombia, Bogotá D.C., Colombia 2011.

[2] Moore, A. W. (2003). Support Vector Machines, 1–33.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of statistical learning* (Second., p. 740). California: Springer Series in Statistics.

[4] Horning, (2011a). Decision Trees & Random Forests

[5] Horning, (2011b). Error analysis in Random Forests.

[6] Introduccion a R: Notas sobre R: Un entorno de programacion para An_lisis de Datos y Gracos. Version 1.0.1 (2000-05-16)

[7] Metodologías inductivas y deductivas en técnicas de teledetección, Mariana Pagot. www.efn.uncor.edu/departamentos/estruct/lgodoy /Metodologia/Documentos/Pagot.pdf.