Análisis de Datos de Pacientes con Riesgo de Presentar CHD usando Técnicas de Machine Learning

Segura-Tinoco, Germán Andrés
Universidad de Los Andes
Bogotá – Colombia
ga.segurat@uniandes.edu.co

Orozco-Cacique, Johanna Alexandra
Universidad de Los Andes
Bogotá – Colombia
ja.orozco@uniandes.edu.co

Julio 2017

ABSTRACT

In the current era in which we live, there is a clear and irreversible tendency to generate and store large volumes of information, from various sources such as: government agencies, public and private companies, clinics and hospitals, social networks, etc. Hence the great need to analyze the data in order to obtain some benefit for the owner thereof, a third party or humanity in general. With this in mind, we conducted a descriptive and predictive analysis of public medical data of South Africa on patients with possible risk of presenting coronary heart disease, and applying advanced techniques of supervised machine learning and models calibration, we were able to determine when a person has high probabilities (close to 70%) of presenting or developing this disease, with the objective of being able to contribute to an early detection and diagnosis of it, for further treatment. Hopeful and convincing results were obtained, which can be improved if there is a greater amount of source data from which to learn.

Palabras claves: Machine Learning, Supervised Learning, Data Analytics, Naive Bayes, SVM, CHD, R.

1. Introducción

La CHD (enfermedad coronaria cardiaca o cardiopatía coronaria) consiste en un estrechamiento de los pequeños vasos sanguíneos que suministran sangre y oxígeno al corazón [1], y es considerada como una de las principales causas de muerte en hombres y mujeres en todo el mundo.

Son varios los factores de riesgo que aumentan la probabilidad de presentar la enfermedad u otra afección cardiaca. Algunos como la edad, el género y los genes, son intrínsecos y no pueden cambiarse, pero otros como si la persona es fumadora, si tiene el colesterol alto, la presión arterial alta, si es diabético o si tiene sobrepeso, entre otros, son factores sensibles que pueden controlarse a partir del cambio de hábitos o del uso de algunos medicamentos [2].

“Hoy en día la prevención de la cardiopatía coronaria se basa prácticamente en dos conceptos significativamente diferentes. La primera es la

educación general de toda la población sobre los factores de riesgo conocidos, mientras que la otra es la prevención basada en la detección de factores de riesgo en la práctica general" [3].

Por lo tanto, es primordial agotar esfuerzos en determinar si una persona tiene probabilidades de presentar la enfermedad o de tener alguna otra afección relacionada. Esto ayuda en varios sentidos, ya que ofrece la oportunidad de mejorar la calidad de vida de las personas que pueden o están desarrollando la enfermedad CHD, además de ser un alivio para los sistemas de salud pública que tienen que invertir óptimamente sus recursos en el tratamiento de enfermedades, que cómo ésta, pueden evitarse.

Por consiguiente, se desarrolló este trabajo, con el objetivo principal es analizar la viabilidad de que a partir de una muestra significativa tomada en hombres de una región del Cabo Occidental de Sudáfrica, se pueda crear un modelo predictivo que ayude a la detección temprana de pacientes que puedan presentar o no la CHD.

Además, se describen los datos utilizados, la aplicación de 6 algoritmos de aprendizaje supervisado de máquinas, la estimación del error global a través de la técnica *Cross-Validation*, la evaluación de los resultados obtenidos, y se presentan unas conclusiones que le permiten al lector entender el proceso predictivo y de generación de conocimiento a partir del uso de estas técnicas.

La implementación de los algoritmos, así como la generación de resultados, se realizó con el lenguaje estadístico R 3.4.1 [4] y el *software* RStudio.

2. Exploración y Descripción de los Datos

Los datos utilizados son una muestra de 462 registros de un *dataset* más amplio, descrito en Rousseauw et al, 1983, South African Medical Journal, perteneciente a una organización sin ánimo de lucro llamada *South African Heart Association* (SAHA), la cual, es una de las más representativas de Sudáfrica en estos temas y es la encargada de "representar y proteger los intereses de profesionales cardiólogos y cirujanos cardíacos, así como el bienestar público mediante la educación dirigida a la prevención y tratamiento de enfermedades del sistema cardiovascular" [5].

Estos datos contienen algunos factores de riesgo de presentar CHD en hombres, incluso muchos de los casos incluidos y registrados como positivos de la enfermedad, "han sido sometidos a tratamiento de reducción de la presión arterial y otros programas para reducir sus factores de riesgo después de su evento de CHD. En algunos casos las mediciones se hicieron después de estos tratamientos" y se registraron 2 controles por cada caso positivo [6].

Las variables independientes o predictoras disponibles en el *dataset* son:

- **sbp:** Presión Sanguínea Sistólica / Systolic Blood Pressure
- **tobacco:** Tabaco acumulado (kg) / Cumulative Tobacco (kg)
- **ldl:** Lipoproteína de Baja Densidad – Colesterol / Low Density Lipoprotein Cholesterol
- **adiposity:** Adiposidad.
- **famhist:** Antecedentes familiares de enfermedad cardíaca (Presente, Ausente) / family history of heart disease (Present, Absent)
- **typea:** Comportamiento de tipo A / type-A behavior
- **obesity:** Obesidad
- **alcohol:** Consumo actual de Alcohol / current alcohol consumption
- **age:** Edad de inicio / age at onset

Para facilitar la posterior aplicación de las técnicas de *Machine Learning*, se decidió trabajar con variables independientes cuantitativas (en vez de cualitativas). Así que, la variable **famihist** inicialmente con valores Presente / Ausente, fue transformada a una de tipo cuantitativa y dicotómica, almacenando los valores 1 como Presente y 0 como Ausente.

La variable dependiente o a predecir es:

- **chd:** Enfermedad Coronaria Cardíaca / coronary heart disease

saha_data

Filter

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	1	49	25.30	97.20	52	Si
2	144	0.01	4.41	28.61	0	55	28.87	2.06	63	Si
3	118	0.08	3.48	32.28	1	52	29.14	3.81	46	No
4	170	7.50	6.41	38.03	1	51	31.99	24.26	58	Si
5	134	13.60	3.50	27.78	1	60	25.99	57.34	49	Si
6	132	6.20	6.47	36.21	1	62	30.77	14.14	45	No
7	142	4.05	3.38	16.20	0	59	20.81	2.62	38	No
8	114	4.08	4.59	14.60	1	62	23.11	6.72	58	Si
9	114	0.00	3.83	19.40	1	49	24.86	2.49	29	No
10	132	0.00	5.80	30.96	1	69	30.11	0.00	53	Si

Showing 1 to 10 of 462 entries

Figura 1: Visualización desde RStudio de los primeros 10 registros del *dataset* usado.

El primer análisis que se realizó, fue el de *Box Plot* [7], para poder tener una mejor comprensión de la distribución de los datos por variable y encontrar posibles atipicidades en los mismos. Como se puede observar en la figura 2, sólo las variables **sbp** y **alcohol** tienen datos atípicos, los cuales no afectaron los posteriores análisis.

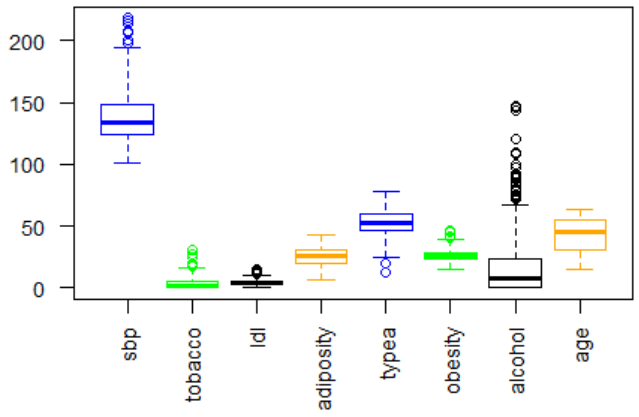

Figura 2: Resultado del análisis univariado de Box Plot o Diagrama de Caja.

El segundo análisis descriptivo de los datos que se realizó, fue el cálculo de correlación entre todas las variables presentes en el *dataset*.

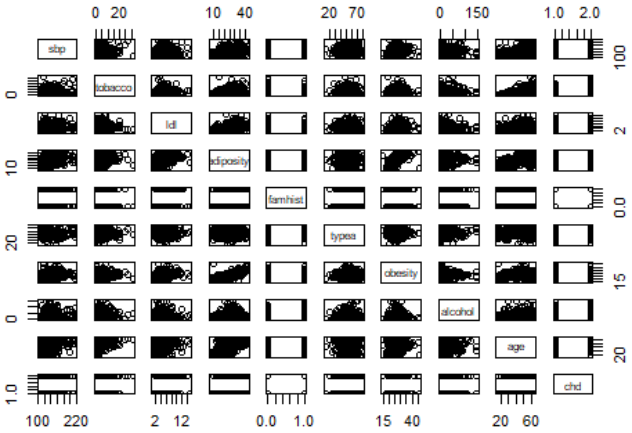

Figura 3: Tabla de correlación entre todas las variables (10x10).

No obstante, encontramos que no hay una correlación bivariada fuerte entre las variables a analizar, además, pareciera que no hay una variable principal que influencie claramente a la variable **chd**.

Así mismo para T. Hastie [8], usando un modelo de regresión logística por máxima verosimilitud, las variables **sbp** y **obesity** no son relevantes dado que su *p-value* supera el nivel de significación del 5%. Sin embargo, cuando se tuvo en cuenta el Criterio de Información Akaike (AIC) para un modelo de supresión gradual de términos por *Splines* Naturales, dichas variables fueron incluidas de nuevo.

Terms	Df	Deviance	AIC	LRT	P-value
none		458.09	502.09		
sbp	4	467.16	503.16	9.076	0.059
tobacco	4	470.48	506.48	12.387	0.015
ldl	4	472.39	508.39	14.307	0.006
famhist	1	479.44	521.44	21.356	0.000
obesity	4	466.24	502.24	8.147	0.086
age	4	481.86	517.86	23.768	0.000

Figura 4: Tabla con el resultado del modelo de regresión logística, después de la supresión gradual de términos de *Splines* Naturales.

Por lo tanto, para nuestro análisis exploratorio y predictivo se decidió usar todas las variables independientes del *dataset*.

Posteriormente, se utilizó la técnica de Análisis de Componentes Principales (PCA) [9], para realizar una exploración de los datos que permitiera determinar la distribución de los mismos, de acuerdo a la varianza y covarianza acumulada en sus componentes.

El PCA nos muestra que en las dos primeras componentes se tiene casi un 45% de la varianza total, es decir, que la información complementaria se encuentra repartida casi equitativamente en todas las otras componentes. Esto sugiere que es un sistema de datos complejo con poca correlación entre sí, donde las variables tienen mucha varianza. Razón por la cual se escalaron los datos, con la finalidad de que las magnitudes de algunas de las variables no opaquen el análisis de las otras.

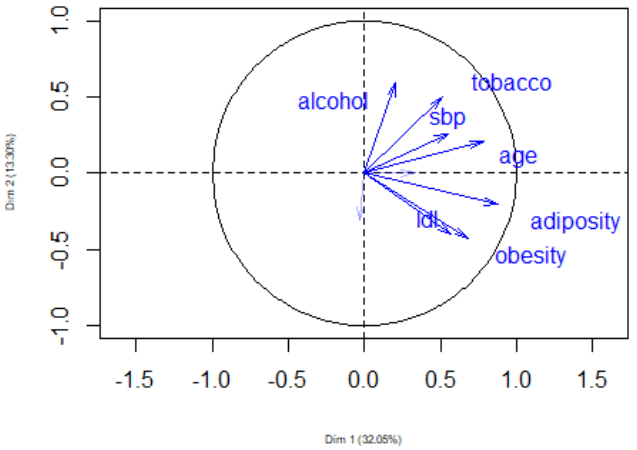

Figura 5: Círculo de correlación, generado a partir del PCA escalado.

Por último, se puede observar en la figura 5, que en general, todas las variables tienen poca correlación entre sí, aunque en todos los casos las correlaciones son positivas.

3. Definición de los Métodos

La selección del algoritmo para resolver un problema en específico es una tarea crítica que requiere tiempo y conocimiento, ya que a menudo se deben tomar en cuenta requisitos no funcionales como son: el tamaño, calidad y naturaleza de los datos, las variables a seleccionar, el objetivo del algoritmo, la implementación, desempeño y rendimiento del mismo, etc.

Por lo tanto, dado nuestro escenario, en donde tenemos en los datos iniciales nueve variables de entrada (independientes) cuantitativas y una de salida (dependiente) cualitativa, se optó por usar técnicas de Aprendizaje Estadístico Supervisado [10] para resolver el problema, enfocando la solución en la predicción de personas que pueden presentar o no CHD (respuesta dicotómica), a través del aprendizaje y clasificación de un conjunto de factores de riesgo registrados para ellos.

En cuanto al objetivo de cada algoritmo, empleamos los 6 que podían generar los mejores resultados dadas las características del problema. A continuación, resumimos cada uno de ellos, para ofrecer una mejor comprensión de los mismos y de sus resultados.

3.1 Support Vector Machine (SVM) Es un algoritmo de clasificación de margen máximo basado en la teoría del aprendizaje estadístico. SVM realiza tareas de clasificación al maximizar el margen que separa ambas clases mientras minimiza los errores de clasificación, a través de un hiperplano [11].

3.2 K-Nearest Neighbors (KNN) Este es un algoritmo clasificador que busca estimar la

distribución condicional de *Y* dado *X*, y luego clasificar una observación dada a la clase con mayor probabilidad estimada. Es decir, dado un participante positivo *K* y una observación de prueba x_0, el clasificador KNN identifica primero los puntos *K* en los datos de entrenamiento que están más cerca de x_0, representados por N_0. A continuación, estima la probabilidad condicional para la clase *j* como la fracción de puntos en N_0 cuyos valores de respuesta son iguales a *j*.

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \qquad (3.1)$$

Finalmente, KNN clasifica la observación de prueba x_0 a la clase con mayor probabilidad [11].

3.3 Naive Bayes Un clasificador de Bayes ingenuo asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. Es especialmente apropiado cuando la dimensión *p* del espacio de características es alta, haciendo la estimación de densidad poco atractiva. El clasificador de Bayes ingenuo asume que dada una clase *G = j*, las características X_k son independientes [8].

$$f_j(X) = \prod_{k=1}^{p} f_{jk}(X_k) \qquad (3.2)$$

3.4 Árboles de Decisión Este algoritmo recursivamente separa las observaciones en las ramas para construir un árbol con el fin de mejorar la exactitud de la predicción. Al hacerlo, utilizan la ganancia de información de algoritmos matemáticos para identificar una variable y un umbral correspondiente para la variable que divide la observación de entrada en dos o más subgrupos. Este paso se repite en cada nodo de la hoja hasta que se construye el árbol completo con cierta profundidad [11].

3.5 Random Forest Los bosques aleatorios construyen una serie de árboles de decisión tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Al construir estos árboles de decisión, cada vez que se considera una división en un árbol, se selecciona una muestra aleatoria de *m* predictores como candidatos del conjunto completo de *p* predictores. Una muestra fresca de *m* predictores se toma en cada división, y típicamente elegimos $m \approx \sqrt{p}$, es decir, el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores. Los bosques aleatorios promueven las divisiones, al considerar sólo un subconjunto de los predictores para cada árbol de decisión [10].

3.6 AdaBoost Es una contracción de "Adaptive Boosting". Funcionalmente busca crear un clasificador fuerte cuya base sea la combinación lineal de clasificadores "débiles simples" $ht(x)$. Sin embargo, AdaBoost propone entrenar una serie de clasificadores débiles de manera iterativa, de modo que cada nuevo clasificador o "weak learner" se enfoque en los datos que fueron erróneamente clasificados por su predecesor, de esta manera el algoritmo se adapta y logra obtener mejores resultados [12].

Con respecto al uso del algoritmo AdaBoost, podemos decir que lo incluimos en nuestro análisis porque la variable a predecir **chd** es dicotómica, lo cual nos permite sacarle provecho a la estructura natural de este método.

4. Configuración de los Modelos

Antes de aplicar cada algoritmo, se utilizó la técnica de validación cruzada (*Cross Validation*), para definir óptimamente los parámetros de ajuste de cada método y así poder obtener la mejor precisión posible al evaluar los resultados de los modelos, reduciendo la posibilidad de caer en el sobreajuste (O*verfitting*).

En la tabla 1 se muestran los parámetros de configuración que se seleccionaron para cada uno de los modelos.

MODELO	*PARÁMETRO*	*VALOR ASIGNADO*
SVM	Kernel	Lineal
K-NN	Kmax	5
Naive Bayes		
Árboles de decisión	Type	Class
Bosques aleatorios	Importance	True
	ntree	300
AdaBoost	Iter	50
	Nu	1
	Type	Real

Tabla 1: Configuración de los parámetros para cada algoritmo usado.

5. Selección del Modelo

Inicialmente, se seleccionaron las métricas con las cuales se evaluarían los modelos, de esta manera se definió que es muy importante predecir la cantidad de positivos para **chd** (porque son los individuos que finalmente deben tener una atención inmediata), junto con la maximización de la precisión global del modelo. Ambas métricas son calculadas a partir de la matriz de confusión.

También, se calculó un error global base de 34,63%, resultado del cociente entre la cantidad de SI = 160 y la cantidad total de registros (462). Éste valor se utilizó como punto de referencia para comparar los resultados de cada modelo.

A continuación, se muestran los resultados de los modelos, luego de ser validados con *Cross Validation* 10 veces, y promediar sus errores globales y locales en cada iteración. De forma empírica, se definió en 10 el número de *folds* para la validación, y se repitió la técnica 10 veces para los 6 algoritmos mencionados anteriormente, con la finalidad de que compitieran lo más justamente entre sí y poder encontrar el que realmente ofreciera la mejor solución.

Como se puede observar en la figura 6, los modelos generados a partir de los algoritmos SVM y Naive Bayes son los que tienen el menor error global promedio en sus predicciones, con valores claramente menores al error global base de 34,63%.

Si se desea calcular la precisión global del modelo (para tener otro punto de vista), se puede usar la siguiente fórmula: [Precisión Global] = 1 – [Error Global].

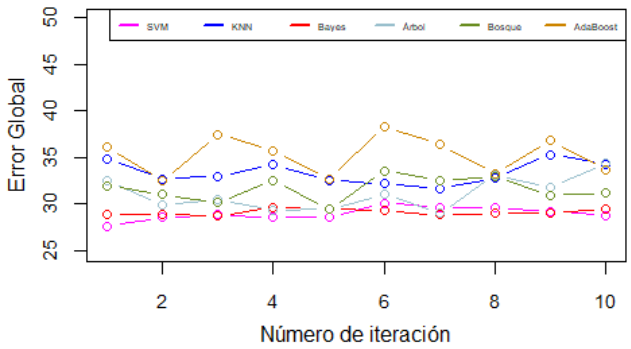

Figura 6: Error global promedio de cada modelo, en cada una de las 10 pruebas realizadas.

En cuanto a la cantidad de SI detectados por cada modelo, Naive Bayes es el claro ganador, con casi 100 predicciones correctas de 160 (figura 7). Y al revisar la cantidad de NO detectados en la figura 8, vemos que el mejor desempeño es para los modelos de SVM y Bosques Aleatorios. Sin embargo, ésta métrica no es tan relevante, como las 2 anteriores.

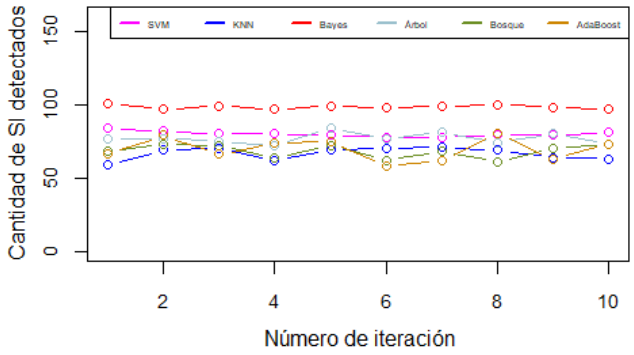

Figura 7: Cantidad de SI detectados por cada modelo, en cada una de las 10 pruebas realizadas.

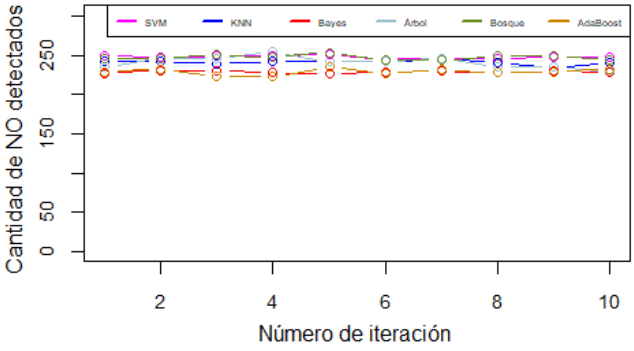

Figura 8: Cantidad de NO detectados por cada modelo, en cada una de las 10 pruebas realizadas.

MODELO	*%GLOBAL ERROR*	*% SI DETECTED*	*% NO DETECTED*
SVM	28.71	50.00	82.45
Naive Bayes	28.77	61.88	76.16
Árboles de Decisión	31.22	48.75	79.47
Bosques Aleatorios	31.77	41.88	82.12
K-NN	33.08	41.88	80.46
Error Base	34.63	NA	NA
AdaBoost	34.95	43.13	76.82

Tabla 2: Resumen de la evaluación de los 6 algoritmos seleccionados.

Basados en la información anterior, seleccionamos al modelo Naive Bayes como mejor predictor para el problema que estamos tratando, porque tiene un error global promedio excelente de 28.77% y es el que tiene el mayor índice de detección de SI (indicador clave), con 61.88%.

Otro de los factores que influyó en la selección del modelo Naive Bayes, es que éste no realiza *Overfitting* a ninguna de las 2 clases a predecir, lo cual lo vuelve un modelo más robusto, que puede generalizar mejor en datos nuevos.

6. Comparación contra otros Modelos

Con la finalidad de confirmar la selección del modelo más eficiente, contrastamos nuestros resultados con los obtenidos por otros autores que usaron los mismos datos.

Por ejemplo, el modelo de Mezclas para Estimación de Densidad, usado en el libro de Hastie [8], encuentra dos subpoblaciones para la variable CHD con la siguiente distribución:

		Mixture model	
		$\hat{\delta} = 0$	$\hat{\delta} = 1$
CHD	No	232	70
	Yes	76	84

Tabla 3: Matriz de confusión para el modelo de Mezclas para Estimación de Densidad.

Con dicha información se obtiene que el error global es del 31.60%, el cual es mayor al obtenido por nuestro modelo con Naive Bayes del 28.77%.

7. Conclusiones

Sí es posible crear un modelo estadístico de aprendizaje, que ayude en la detección temprana de pacientes con tendencia a presentar enfermedades coronarias cardiacas; en el caso de Naive Bayes, con una precisión del 61.88%.

Éste modelo no reemplazaría en ningún momento la opinión experta de un cardiólogo, sin embargo, puede ser muy útil para preseleccionar candidatos a padecer CHD.

Los algoritmos estadísticos de aprendizaje maximizan sus resultados, a medida que existen más datos balanceados a partir de los cuales aprender, por lo tanto, es muy probable que la precisión de los algoritmos usados en éste trabajo mejore, si se utiliza un *dataset* más completo.

Reconocimiento

Al PhD. Oldermar Rodriguez, por todo su conocimiento compartido en los campos de *Machine Learning* y *Data Mining*.

Referencias

[1]. A.D.A.M, Inc. (14 de Julio de 2015). *U.S. National Library of Medicine*. Recuperado el 1 de Julio de 2017, de Medline Plus - Información de Salud para Usted: https://medlineplus.gov/spanish/ency/article/007115.htm

[2]. A.D.A.M, Inc. (2 de Agosto de 2016). *U.S. National Library of Medicine*. Recuperado el 1 de Julio de 2017, de Medline Plus - Información de salud para usted: https://medlineplus.gov/spanish/ency/patientinstructions/000106.htm

[3]. Krstacic, G., Gamberger, D., & Smuc, T. (2001). Coronary Heart Disease Patient Models Based on Inductive Machine Learning. *Artificial Intelligence in Medicine*, 113–116.

[4]. © The R Foundation. (s.f.). *R*. Recuperado el 6 de Julio de 2017, de The R Project for Statistical Computing: https://www.r-project.org/

[5]. South African Heart Association NPC. (s.f.). *South African Heart Association*. Recuperado el 1 de Julio de 2017, de SA Heart®: https://www.saheart.org/

[6]. Stanford University - Department of Statistics. (s.f.). *Department of Statistics*. Recuperado el 27 de Junio de 2017: https://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.info.txt

[7]. Fundación Wikimedia, Inc. (29 de Abril de 2017). *Wikipedia®*. Recuperado el 8 de Julio de 2017, de Wikipedia - La enciclopedia libre: https://es.wikipedia.org/wiki/Diagrama_de_caja

[8]. Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Stanford: Springer.

[9]. Fundación Wikimedia, Inc. (8 de Julio de 2017). *Wikipedia®*. Recuperado el 9 de Julio de 2017, de Wikipedia - La enciclopedia libre: https://en.wikipedia.org/wiki/Principal_component_analysis

[10]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

[11]. Chen, J., Xing, Y., Xi, G., Chen, J., Yi, J., Zhao, D., & Wang, J. (2007). A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease. *Advances in Neural Networks - ISNN 2007*, 1274-1279.

[12]. Morales Sánchez, A. A. (Mayo de 2015). *Capítulo 3. Clasificadores Débiles - AdaBoost.* Recuperado el 1 de Julio de 2017, de Colección de Tesis Digitales - Universidad de la Ámericas Puebla: http://catarina.udlap.mx/u_dl_a/tales/documentos/lmt/morales_s_aa/