

**DISEÑO DE UN MODELO PARA ASISTENCIA MÉDICA Y
PREDICCIÓN DE PREECLAMPSIA EN MUJERES EMBARAZADAS**

ROBERTO STEVENS PORTO SOLANO
Ingeniero de Sistemas, Candidato a Magíster en Ingeniería de Sistemas

Trabajo final de investigación presentado como requisito para optar por el título de Magíster
en Ingeniería de Sistemas.

DIRECTOR
ING. MIGUEL JIMENO PH.D

FUNDACIÓN UNIVERSIDAD DEL NORTE
DIVISIÓN DE POSGRADOS
MAESTRIA EN INGENIERIA DE SISTEMAS
BARRANQUILLA, NOVIEMBRE DE 2015

Nota de Aceptación

Director

Evaluador

Evaluador

AGRADECIMIENTOS

Quiero aprovechar estas líneas para expresar mi agradecimiento a quienes han hecho posible que esta tesis de Maestría fuera posible.

A Dios, nada en mi vida sería posible sin ti padre.

A mi esposa, por su apoyo fundamental en esta nueva etapa de mi vida.

A mi hijo, por ser el impulso de mi diario vivir.

A mis padres, quienes siempre me apoyaron incondicionalmente y nunca dudaron que fuera posible.

A mi Director de tesis, Miguel Jimeno, por su continuo apoyo, orientación y consejo en cada etapa de este proceso.

Al Director médico de la clínica Portoazul, Nicola Ambrossi, quien me brindo el apoyo y la oportunidad de poder desarrollar esta investigación.

A mi hermano, por su apoyo incondicional y consejos.

A mi tía, la Doctora Katherine Porto, por su continua asesoría a lo largo de este proceso.

A mi amigo Danilo García por su apoyo fundamental y consejos en este proceso.

A los profesores Jesús Cohen, Jorge Oyola y Harold Pérez, por su apoyo.

A todos los que me educaron y estuvieron pendientes en mi proceso de formación.

TABLA DE CONTENIDO

1.	INTRODUCCIÓN.....	8
1.1.	OBJETIVOS	9
1.1.1.	<i>General</i>	9
1.1.2.	<i>Específicos</i>	9
1.1.	MOTIVACIÓN.....	10
1.2.	ORGANIZACIÓN DEL DOCUMENTO	10
2.	ESTADO DEL ARTE	11
2.1.	MARCO TEÓRICO	15
2.1.1.	<i>Técnicas para Predicción y Minería de Datos</i>	15
2.1.2.	<i>Distribuciones Estadísticas</i>	34
2.1.3.	<i>Análisis De La Varianza Con Un Factor (Anova)</i>	44
2.1.4.	<i>Permanova</i>	45
2.1.5.	<i>Pruebas No Paramétricas</i>	45
2.1.6.	<i>Pruebas T</i>	48
2.2.	TRASTORNOS HIPERTENSIVOS EN LAS MUJERES EMBARAZADAS (NORMALES)	50
2.3.	FACTORES A DIAGNOSTICAR EN LA PREECLAMPSIA	52
2.3.1.	<i>Factores de Riesgo</i>	55
2.4.	EFFECTOS DE LA PREECLAMPSIA	56
2.5.	TRATAMIENTO DE LA PREECLAMPSIA GRAVE.....	57
2.6.	TRATAMIENTO DE LA PREECLAMPSIA LEVE.....	58
3.	ASPECTOS METODOLÓGICOS.....	60
3.1.	TIPO Y DISEÑO DE LA INVESTIGACIÓN	60
3.2.	ÁREA DE ESTUDIO.....	61
3.3.	UNIVERSO	61
3.4.	MUESTRA.....	61
3.5.	CRITERIOS DE INCLUSIÓN.....	61
3.6.	CRITERIOS DE EXCLUSIÓN	62
3.7.	INSTRUMENTO DE RECOLECCIÓN DE LA INFORMACIÓN.....	62
3.8.	DESCRIPCIÓN DE PROCEDIMIENTOS	62
4.	MODELO DE PREDICCIÓN DE PREECLAMPSIA.....	63
4.1.	MODELOS DE EXTRACCIÓN DE CONOCIMIENTO PARA LA PREDICCIÓN DE PREECLAMPSIA	63
4.2.	PREPARACIÓN Y SELECCIÓN DE DATOS	65
4.3.	ALGORITMO DE EXTRACCIÓN	67
5.	IMPLEMENTACIÓN DEL MODELO	68
5.1.	ALGORITMO C5.0 EN R	68
5.2.	SISTEMA DE PREDICCIÓN NMDS EN R	73
5.2.1.	<i>Principal Component Analysis PCA</i>	73
5.2.2.	<i>Principal Coordinate Analysis PCoA o Non Metric Dimensional Scaling</i>	77
6.	PRUEBAS Y RESULTADOS	81
7.	CONCLUSIONES Y PERSPECTIVA	84
	ANEXOS.....	86
	BIBLIOGRAFÍA.....	92

LISTA DE FIGURAS

FIGURA 1. REPRESENTACIÓN LINEAL DE DATOS	15
FIGURA 2. EJEMPLO DE ÁRBOL BINARIO	27
FIGURA 3. ALGORITMO DIAGNÓSTICO PARA LA CLASIFICACIÓN DE TRANSTORNOS IMPERTENSIVOS.....	51
FIGURA 4. CLASIFICACIÓN DE PREECLAMPSIA	51
FIGURA 5. MODELO DE EXTRACCIÓN DE CONOCIMIENTO PARA LA DETECCIÓN DE PREECLAMPSIA	65
FIGURA 6. DATASET PREECLAMPSIA	68
FIGURA 7. ENCABEZADO DEL DATASET	69
FIGURA 8. ESTRUCTURA DEL DATASET	69
FIGURA 9. ARBOL ENTRENADO C5.0	73
FIGURA 10. CORRELACIÓN ENTRE VARIABLES	74
FIGURA 11. DATASET EN FUNCION DEL PCA.....	75
FIGURA 12. SEPARACIÓN FACTORES PREDICTORES	76
FIGURA 13. DISTANCIA ENTRE FACTORES	79
FIGURA 14. DISTANCIA ENTRE LAS ESPECIES DE UN DATASET	79
FIGURA 15. MEDIDA DE AJUSTE PARA PUNTOS DE ESCALAMIENTO MULTIDIMENSIONAL.....	80
FIGURA 16. ARBOL C5.0 - 2 NIVEL	82
FIGURA 17. RESULTADOS DEL ENTRENAMIENTO C5.0	82

Resumen

El presente proyecto presenta un modelo basado en un sistema de información para la asistencia de las mujeres embarazadas el cual les ayudara a tener un mejor control de su embarazo y para el médico poder estar mejor informado sobre las probabilidades de que sus pacientes puedan padecer de preeclampsia. Dentro del proyecto se desarrollo un modelo de predicción de preeclampsia en mujeres embarazadas durante el segundo trimestre de gestación, empleando técnicas de minería de datos (como lo son los arboles de decisión, Algoritmo C5.0, Principal Component Analysis PCA, Non Metric Dimensional Scaling NMDS). La propuesta de este modelo surge del interés de diseñar una herramienta de predicción de preeclampsia, con el fin de ayudar con los expertos en obstetricia a examinar y verificar más fácilmente los resultados obtenidos de sus pacientes para apoyar la toma de decisiones. Una vez se determinaron las variables más acordes para hacer la predicción, se procedió a seleccionar el mejor método para lograr este proceso. El modelo fue probado sobre un conjunto de 100 datos simulados de los cuales el 75% de la muestra, corresponde a datos de entrenamiento y el 25 %, a datos de validación. Los resultados obtenidos de la técnica utilizada (C5.0, PCA, NMDS) se compararon con otras técnicas de predicción encontradas en revisión literaria mejorando el porcentaje de acierto.

Palabras Claves: Predicción de Preeclampsia, técnicas de minería de datos, arboles de decisión, Non Metric Dimensional Scaling NMDS, Principal Component Analysis PCA, Algoritmo C5.0, Análisis multivariado.

Abstract

This project is based on an information system for the care of pregnant women which will help them take better control of their pregnancy and for health to be better informed about the chances that their patients can suffer from preeclampsia model . Within the project a model prediction of preeclampsia in pregnant women during the second trimester of pregnancy, using data mining techniques (such as decision trees, C5.0 algorithm, Principal Component Analysis PCA, Non Metric was developed Dimensional Scaling NMDS). The proposal of this model stems from the interest of designing a tool for predicting preeclampsia, in order to help experts in obstetrics to more easily examine and verify the results of their patients to support decision-making. Once the most consistent variables were determined to make the prediction, we proceeded to select the best method to achieve this process. The model was tested on a simulated data set of 100 of which 75% of the sample corresponds to training data and 25%, data validation. The results of the technique used (C5.0, PCA, NMDS) were compared with other predictive techniques found in literature review to improve the percentage of success.

Keywords: Preeclampsia prediction techniques, data mining, decision trees, Non Metric Dimensional Scaling NMDS, Principal Component Analysis PCA, C5.0 algorithm, multivariate analysis.

CAPÍTULO 1

1. INTRODUCCIÓN

La preeclampsia es el trastorno hipertensivo más frecuente en el embarazo y constituye la primera causa de mortalidad materna en nuestro medio (Cifuentes R., 2006). La incidencia mundial es de 2-10% (Harskamp RE., 2007). En E.U.A. afecta al 3-8% de las mujeres en embarazo, es la segunda causa de muerte materna luego del embolismo representando aproximadamente un 15 % de esas muertes (Sibai B, 2005). Así mismo es la primera causa de prematuridad iatrogénica en los países industrializados. En Colombia afecta aproximadamente al 10-15% de la población (Cifuentes, 2010).

El origen de la preeclampsia es desconocida. Se sabe que hay un daño endotelial ocasionado por falla en la segunda etapa de migración trofoblástica, lo cual genera una disminución en la producción de sustancias vasodilatadoras (óxido nítrico, Prostaciclina) con aumentos de sustancias vasoconstrictoras (endotelina, tromboxano y angiotensina II) resultando una agregación plaquetaria, alteración en la reactivación cardiovascular e hipertensión arterial.

Más recientemente se ha encontrado que la disfunción endotelial de la preeclampsia se asocia también a un incremento en la producción de factores antiangiogénicos: sFlt-1 (forma soluble de tirosina kinasa) y sEng (endoglina Soluble). Estos antagonizan los factores angiogénicos: VEGF (factor de crecimiento del endotelio vascular) y PlGF (factor de crecimiento placentario), creándose un balance negativo de estos últimos. En general se considera que una paciente es

hipertensa en el embarazo cuando tiene una presión arterial igual o mayor que 140 y /o 90mmHg (Cifuenes R, 2009).

El problema de la detección de preeclampsia radica en el análisis de síntomas que permitan conocer el comportamiento (estado) de un proceso gestacional, con el fin de detectar anomalías. Según (Vázquez, 2011) a menudo las mujeres que padecen preeclampsia no se sienten enfermas, ya que algunos síntomas iniciales en el embarazo suelen confundirse con trastornos normales del embarazo.

1.1. Objetivos

1.1.1. General

Diseñar e implementar un modelo para asistencia médica y predicción de preeclampsia en mujeres embarazadas.

1.1.2. Específicos

1. Determinar y analizar factores causales de preeclampsia en mujeres embarazadas durante el segundo trimestre de gestación.
2. Diseñar e implementar un modelo de predicción de factores causales de preeclampsia.
3. Validar el modelo expuesto de predicción de factores de preeclampsia con un conjunto de datos simulados.
4. Diseñar e implementar un software para asistir al médico a realizar predicciones más acertadas durante el proceso de gestación de la paciente.
5. Evaluar la eficiencia del modelo de predicción al compararse con otros métodos actualmente usados.

1.1. Motivación

La finalidad de esta tesis está basada en la realización de un modelo de apoyo tanto para las mujeres embarazadas como para el médico que las acompaña en su etapa de gestación. Se pretende reducir el riesgo que sufren las mujeres embarazadas en el último trimestre de presentar la enfermedad que las pueda conducir a una eclampsia produciendo la muerte de la paciente. Para el médico el poder tener una herramienta que le permita dar predicciones más acertadas para la prevención de preeclampsia.

1.2. Organización del Documento

Este documento está diseñado de tal forma que los conceptos aplicados no se presentan a detalle, sino en forma general, en su lugar se hacen referencias a la literatura que presenta una descripción más compleja de dichos conceptos. Esta tesis está organizada en:

Capítulo 2. Contiene las bases teóricas que soporta el trabajo presentado. Este comprende, técnicas de minería de datos, técnicas de predicción de preeclampsia, además del algoritmo de entrenamiento utilizado, resumen del capítulo.

Capítulo 3. Se describe la metodología utilizada en la investigación

Capítulo 4. Describe el modelo propuesto predicción de preeclampsia. Para ello se describe el algoritmo de árboles de decisión propuesto al igual que la técnica de predicción seleccionada para el modelo.

Capítulo 5. Presenta los resultados del modelo desarrollado, el cual fue probado sobre un conjunto de datos obtenidos de cuadros clínicos de mujeres embarazadas de una Sociedad Anónima Simplificada de Colombia, para ello se adelantó varias pruebas con diferentes variables para el entrenamiento de la red neuronal.

Capítulo 6. Pruebas y resultados del modelo

Capítulo 7. Se presenta conclusiones del trabajo realizado y algunas ideas para trabajos futuros.

CAPÍTULO 2

2. ESTADO DEL ARTE

En este capítulo se establecen las bases teóricas del proyecto al igual que el conocimiento técnico previo que hay que tener en cuenta para el desarrollo de este; con la finalidad de hacer más fácil la lectura del capítulo. El capítulo comienza planteando el problema de predicción de preeclampsia, luego se presenta las principales técnicas utilizadas para la detección de preeclampsia, comenzando con las técnicas tradicionales y posteriormente se presentan las técnicas de minería de datos las cuales han permitido encontrar nuevos patrones en base a la información almacenada. Posteriormente, se exponen las características de árboles de decisión, redes neuronales, regresión lineal y se presenta como técnica de solución del problema de predicción de preeclampsia a partir de unas reglas de entrenamiento definidas por los algoritmos C5.0, PCA, NMDS, Cart que permiten una mejor comprensión de los resultados obtenidos por parte de los expertos en el tema. Finalmente se presentan las técnicas más usadas para la predicción.

Según (Bautista, 2012) los trastornos hipertensivos asociados al embarazo son muy frecuentes durante el control prenatal. En nuestro medio es una entidad endémica presente hasta en 10% de los embarazos. En la práctica diaria esta prevalencia tan importante se acompaña de gran morbilidad y mortalidad materna y perinatal. La primera causa de muerte materna en nuestro país está relacionada con la toxemia gravídica. La entidad obstétrica más relacionada con estos casos letales es la eclampsia, acompañada o agravada por el Síndrome de Hemolytic anemia Elevated Liver Enzyme Low Platelet HELLP. La manifestación hipertensiva más frecuentemente encontrada durante el embarazo es la preeclampsia. Esta entidad aparece en gestaciones menores de 34 semanas en 35% de los casos. Cuando se encuentra la asociación de preeclampsia y prematuridad se

conjuga el determinante clínico de 30% de las muertes perinatales en nuestro medio. Es así como la preeclampsia es la segunda causa de muerte perinatal. A nivel mundial se ha intentado controlar de manera preventiva la aparición de los cuadros clínicos severos relacionados a la hipertensión arterial durante la gestación sin lograr el éxito esperado. Estas entidades están muy relacionadas con las condiciones socioeconómicas de la población, lo cual hace imposible su desaparición del escenario clínico. Su frecuencia se dispara en casos de madres solteras, embarazos no deseados, desempleados y desplazados, toda ésta una problemática muy nuestra. Sin embargo, el pronóstico es susceptible de modularse mejorando el diagnóstico temprano.

Según la investigación realizada en el centro de investigación Harris Birthright de Medicina Fetal en Londres por (Neocleous, 2009). De 6838 casos de mujeres embarazadas se registraron por paciente 24 parámetros. De los cuales 15 parámetros fueron considerados, como los de más influencia en la caracterización del riesgo de aparición de preeclampsia. Para la predicción de la preeclampsia fue utilizado una red neuronal feedforward tanto multicapa como multi-losa. En el entrenamiento de la red neuronal lograron una clasificación de los casos del 83,6% de la preeclampsia y en una prueba ajustada de un 93,8%. Los casos de predicción de preeclampsia cuando apareció información de tabaquismo y alcoholismo llegaron las pruebas a un 100%.

Según (Dustin T. Dunsmuir, 2014), el modelo fullPIERS (Preeclampsia Integrate Estimate of RiSk) fue desarrollado para evitar la necesidad de no usar pruebas de laboratorio que con frecuencia son realizadas en hospitales de tercer nivel. Para la efectividad de predicción se toman variables como: simple demografía, los síntomas y signos clínicos; Este modelo de predicción simplificado incluye la paridad, la edad gestacional en la presentación, la presión arterial sistólica, diastólica, la varilla de medición proteinuria, y la presencia de los siguientes síntomas: dolor en el pecho, falta de aliento, dolor de cabeza, trastornos visuales, y sangrado vaginal con dolor abdominal. El desarrollo de la aplicación móvil para el diagnóstico y la gestión de las mujeres embarazadas con preeclampsia es

descrito. Esta aplicación está diseñada para su uso por basado en la comunidad proveedores de atención médica (c-SCPH) en los centros de salud y durante visitas a domicilio para recoger síntomas y realizar mediciones clínicas (incluyendo lecturas del oxímetro de pulso). Los datos clínicos recogidos en las mujeres con preeclampsia se utilizan como entradas a un modelo predictivo, que proporciona una puntuación de riesgo para el desarrollo de resultados adversos. Sobre la base de este riesgo, la aplicación proporciona recomendaciones sobre el tratamiento, remisión, y la reevaluación. c-HCP puede acceder registros de los pacientes a través de múltiples visitas, utilizando múltiples dispositivos que se sincronizan mediante una captura de Investigación electrónico de datos segura servidor. Una característica única de esta aplicación es la capacidad de medir la saturación de oxígeno con un pulsioxímetro conectado a un teléfono inteligente (oxímetro de teléfono). El modelo miniPIERS emite una puntuación, una calificación de probabilidad, la cual es utilizada dentro de un modelo de árbol de decisión para dar las recomendaciones a los tratamientos de los pacientes. Además, el árbol de decisión incluye información de los eventos maternos adversos en el pasado, con la actual etapa del embarazo, y SpO₂.

En (S.Y. Leemaqz, 2013) obtienen los valores influyentes de los exámenes (NPV ,PPV ,r ,s) para la predicción de preeclampsia antes de la semana 15 y dentro de la semana 15 a la 20. Donde **NPV** es la proporción de verdad negativos en los casos previstos de embarazo sin complicaciones, **PPV** es la proporción de verdad positivos en los casos previstos de PE, **r** es la proporción de casos verdaderamente predichos de PE y **s** es la proporción de verdaderamente casos de embarazo sin complicaciones predicho. El modelo expuesto utiliza un algoritmo N.Bayes para predecir la preeclampsia basado en los valores NPV, PPV, r, s obteniendo un porcentaje de aciertos de 81%.

Según (Etchegaray Adolfo, 2012) las características maternas más influyentes para la predicción de preeclampsia se pueden obtener entre las semanas 11 y 13 o sea en el segundo trimestre de gestación. La técnica para la predicción es basada en la definida por el grupo de la Fetal Medicine

Foundation de reino unido, en donde hacen una combinación de factores de riesgos maternos como (IMC, grupo étnico antecedentes de preeclampsia, de presión arterial, combinación de analitos séricos, índice de pulsatilidad de arterias uterinas, proteínas asociadas al embarazo y factor de crecimiento placentario).

Según (Garovic, 2014) la utilización de mediciones de orina y sueros circulantes de marcadores angiogénicos no proporcionan una forma fiable para la detección de preeclampsia. En un intento por mejorar los valores predictivos de los marcadores se combinaron índices angiogénicos como sFlt-1/PlGF y encontraron cambios que demuestran que no es efectiva en la predicción de la preeclampsia de inicio tardío, pero que es útil en la predicción temprana de la enfermedad. Estos resultados plantean la posibilidad de utilizar estas pruebas en los subgrupos de pacientes especiales, tales como aquellos con hipertensión crónica, que puede estar en un nivel superior de riesgo de preeclampsia y las complicaciones relacionadas con preeclampsia.

También (Verlohren, 2012), explica que en numerosos estudios que se han puesto de manifiesto, se denota el papel clave de la placenta en la patogénesis de la preeclampsia. Un cambio en la sFlt-1 (similar a Fms solubles tirosina quinasa-1) / PlGF (factor de crecimiento placentario) relación se asocia con la enfermedad. Aunque la preeclampsia parece ser una enfermedad claramente definida, la presentación clínica, y en particular la dinámica de la evolución clínica, pueden variar enormemente. Las únicas herramientas disponibles a diagnosticar la preeclampsia son la medición de la presión arterial y el muestreo de proteína en la orina. Sin embargo, estas herramientas tienen una baja sensibilidad y especificidad con respecto a la predicción del curso de la enfermedad o los resultados maternos y perinatales. La única cura para la enfermedad es la entrega, aunque un diagnóstico oportuno ayuda en la disminución de la morbilidad y mortalidad materna y fetal. Los índices angionénicos sFlt1 / PlGF en proporción son capaces de dar una valiosa información

adicional sobre el estado y evolución de la enfermedad y es apto para ser implementado en el algoritmo diagnóstico de la preeclampsia.

2.1. Marco Teórico

2.1.1. Técnicas para Predicción y Minería de Datos

2.1.1.1. Regresiones Lineales

El algoritmo de regresión lineal, es una variación del algoritmo de árboles de decisión que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente.

La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos. Por ejemplo, la línea del siguiente diagrama muestra la mejor representación lineal de los datos.

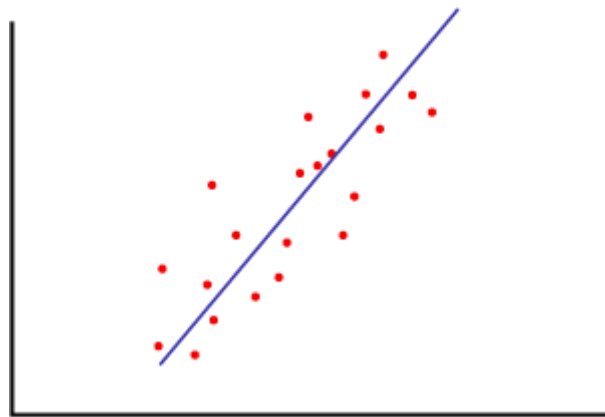


Figura 1. Representación Lineal de Datos

Cada punto de datos del diagrama tiene un error asociado con su distancia con respecto a la línea de regresión. Los coeficientes a y b de la ecuación de regresión ajustan el ángulo y la ubicación de la recta de regresión. Puede obtener la ecuación de regresión ajustando a y b hasta que la suma de los errores asociados a todos los puntos alcance su valor mínimo.

Hay otros tipos de regresión que utilizan varias variables y también hay métodos no lineales de regresión. Sin embargo, la regresión lineal es un método útil y conocido para modelar una respuesta a un cambio de algún factor subyacente.

El modelo lineal relaciona la variable dependiente Y con K variables explícitas X_k ($k = 1, \dots, K$), o cualquier transformación de éstas que generen un hiperplano de parámetros β_k desconocidos:

$$Y = \sum \beta_k X_k + \varepsilon \quad (1)$$

Donde ε es la perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. En el caso más sencillo, con una sola variable explícita, el hiperplano es una recta:

$$Y = \beta_1 + \beta_2 X_2 + \varepsilon \quad (2)$$

El problema de la regresión consiste en elegir unos valores determinados para los parámetros desconocidos β_k , de modo que la ecuación quede completamente especificada. Para ello se necesita un conjunto de observaciones. En una observación i -ésima ($i = 1, \dots, I$) cualquiera, se registra el comportamiento simultáneo de la variable dependiente y las variables explícitas (las perturbaciones aleatorias se suponen no observables).

$$Y_i = \sum \beta_k X_{ki} + \varepsilon_i \quad (3)$$

Los valores escogidos como estimadores de los parámetros $\hat{\beta}_k$, son los coeficientes de regresión sin que se pueda garantizar que coincidan con parámetros reales del proceso generador. Por tanto, en

$$Y_i = \sum \hat{\beta}_k X_{ki} + \hat{\varepsilon}_i \quad (4)$$

Los valores $\hat{\varepsilon}_i$ son por su parte estimaciones o errores de la perturbación aleatoria.

2.1.1.2. Redes Neuronales

El algoritmo de red neuronal combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada.

2.1.1.3. Árboles de Decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta la cual en últimas es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión.

Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión. El árbol de decisión suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un test sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, este tipo de nodos es redondo, los demás son cuadrados. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada (Rokach, 2008).

2.1.1.4. Recursive partitioning

Es un método estadístico multi-variable. Este posicionamiento recursivo crea un árbol de decisión que se esfuerza para clasificar correctamente miembros de población mediante subdivisión de poblaciones en base a variables independientes dicotómicas. El proceso se denomina recursivo si es posible que poblaciones subdivididas pueden dividirse a su vez en un número indefinido de veces. Este proceso de división termina cuando un criterio de parada particular es alcanzado.

Estos métodos han sido desarrollados desde 1980. De los cuales los más conocidos son; ID3, C4.5, C5.0 y Cart. Los cuales ayudan a la clasificación, toma de decisiones en grandes cantidades de datos a través de un modelo de árbol (Cook EF, 1984).

2.1.1.5. ID3

Es un algoritmo usado para generar árboles de decisión. Este algoritmo usa un proceso recursivo sobre todas las ramas del árbol generado. Estas ramas son llamadas así mismo conjunto de ejemplo, el de atributos y el nodo raíz del árbol que estaría vacío. Este proceso se realiza hasta que todos los ejemplos de la rama en cuestión pertenezcan a una única clase. Sin embargo, esto no siempre es posible. En primer lugar, puede ocurrir que se hayan agotado todos los atributos y, sin embargo, sigan existiendo ejemplos con distintos valores de clase. De otra parte, también puede suceder que, una vez elegido un atributo para un nodo de decisión, no exista ningún ejemplo para una de las ramas generadas por dicho atributo. En esos casos, se etiqueta el nodo hoja con la clase mayoritaria (Quinlan, 1986).

Según (Ichihashi, 1995), ID3 adopta una información de base probabilística para inducir los árboles de decisión. Para la adquisición de las normas de producción en los sistemas expertos, tanto en el enfoque de la entrevista y el aprendizaje a partir de ejemplos son indispensables.

(Chi, 1996), explica que el algoritmo es de naturaleza simbólica y no puede modelar dominios en los que hay una gran número de valores continuos como lo hace el algoritmo de Backpropagation. Por lo tanto el ID3 es particularmente útil para producir un conjunto de reglas de clasificación. Luego (Bagchi, 1997), propone un esquema simple pero eficaz para la clasificación de características RID3. En este construye un árbol preliminar con un umbral predeterminado en cada nodo. Si el rendimiento inicial del árbol no es satisfactorio, entonces el umbral en cada nodo estará sintonizado con algoritmos genéticos. Clasificando al vecino más cercano para todos los conjuntos de datos considerados.

En comparación con las redes neuronales, las reglas de extracción son relativamente mucho más fácil de realizarlas con ID3. Esta reduce la entropía en la construcción de un árbol de decisión, dando una ganancia de información y partición de los datos que ayudan a la predicción de la decisión de un resultado. La combinación del algoritmo de las redes neuronales con ID3 sugiere métodos para analizar juicios heurísticos expertos (Mak, 2000).

Para (Shao, 2000), la aplicación de algoritmo ID3 en la adquisición de conocimientos relacionados con la tolerancia, proporciona el cálculo de la medida de ganancia de información que genera el árbol de decisión final, puesto que este examina todos los atributos, eligiendo el de mayor entropía.

(Jin, 2012), propone un algoritmo generalizado de partición difusa ID3 para la heurística de selección de atributos extendidos de un árbol de decisión difusa. Teniendo un gran impacto en la selección de las características no lineales del grado de pertenencia de los conjuntos borrosos.

(Xiaohu, 2012), propone la utilización del algoritmo ID3 para la clasificación de información en páginas comerciales en las cuales se miden como variables la publicidad, datos de compras, pago dirigido a los mercaderes, comportamiento de compra de los usuarios, datos del usuario entre otros.

Construyendo un árbol de decisiones basado en el aumento de la información generando reglas de comportamiento de compras útiles.

Según (Kale, 2015), aplicando el algoritmo ID3 en los datos de alimentos que más consumen los niños es posible seleccionar el mejor alimento, puesto que esto divide la información en los atributos de la muestra y toma decisiones. El árbol de decisión, denota que el resultado es positivo y no indica un resultado negativo.

2.1.1.6. C4.5

Es un algoritmo usado para generar arboles de decisión, es una extensión de ID3. El árbol de decisión generado por el C4.5 puede ser usado para la clasificación. La principal mejora sobre el algoritmo la realizó el propio Quinlan. Entre las mejoras que este algoritmo presentaba sobre ID3 se pueden encontrar las siguientes: Tratamiento de los valores continuos de los atributos cuando se desea saber que atributo es mejor y se tiene uno (o varios) que son continuos, por cada uno de los continuos se crean artificialmente nuevos atributos de la siguiente manera. Por cada atributo continuo, se ordenan los valores del atributo en los ejemplos de la tabla, y se especifica la clase a la que pertenecen los ejemplos, eliminando los demás atributos (Quinlan, C4.5 Programs for Machine Learning, 1993).

Otras de las mejoras que realizó (Quinlan, Improved use of continuous attributes in c4.5, 1996), en las mejoras del C4.5 con respecto al ID3, es que el C4.5 crea un umbral y luego divide las lista en aquellos cuyo valor de atributo es superior al umbral y los que son menores o iguales a él.

Para (Polat, 2009), un sistema híbrido basado en el algoritmo C4.5, mejora la precisión de clasificación de patrones en casos de problemas de clasificación multi-clase incluyendo precisión dermatológica, segmentación de imagen, y conjuntos de datos linfográficos.

(Dai, 2014), propone implementar un algoritmo de árbol de decisión típica, C4.5, utilizando el modelo de programación MapReduce que utilizándolo en paralelo y distribuido procesa grandes conjuntos de datos. Los datos de entrada se divide en divisiones con tamaño apropiado, *mapa* procedimiento toma una serie de pares clave / valor, y genera pares clave / valor procesados, que se pasan a un reductor particular mediante cierta función de partición; Luego de que datos se clasificaron, son arrastrados, por el procedimiento *reducir* que itera los valores que son asociados con la etiqueta específica , produciendo cero o más salidas.

Para (Ruggieri, 2002), con base en la evaluación analítica del C4.5 implementa el EC4.5, una versión más eficiente puesto que adopta la mejor entre tres estrategias para el cómputo de los datos de atributos continuos. Todas las estrategias adoptan una búsqueda binaria del umbral en todo el conjunto de entrenamiento a partir desde el umbral local computado en un nodo. La primera estrategia calcula el umbral local utilizando el algoritmo de C4,5, lo que, en particular, ordena los casos por medio de la método Quicksort. La segunda estrategia utiliza también el algoritmo de C4,5, pero adopta un método de recuento tipo. La tercera estrategia calcula el umbral local utilizando una versión de memoria principal del algoritmo de la selva tropical, la cual no necesita la clasificación. Nuestra aplicación calcula la árboles de decisión mismos como C4.5 con una ganancia de rendimiento de hasta 5 veces.

(Agrawal, 2013), resalta que el C4.5 es un bien algoritmo de clasificación pero cuando se utiliza en cálculos de masa, la eficiencia es muy baja. Por los cual propone el C4.5 mejorado con el uso de la regla de L'Hopital lo que simplifica el proceso de cálculo y mejorando la eficiencia en la decisión algoritmo.

Para (Bouchard, 2011), el C4.5 trabaja con un conjunto de datos de entrenamiento para reconocer la actividad, utilizando un conjunto reducido de planes para contener mínimamente los datos de entrenamiento. El C4.5 utiliza estos registros de entrenamiento para la conclusión de un más plausible. Este solo utiliza un conjunto de entrenamiento contenido, en donde se deben ignorar todos los datos de formación en relación con un plan que no figura en nuestras hipótesis. Por lo cual se construye un árbol de decisión que va a utilizar los datos restantes con el fin de decidir qué actividad del agente creará estar en curso.

(Mantas, 2014), propone una modificación del algoritmo C4.5 en donde se realice un proceso de poda para resolver el problema de over-fitting. Este nuevo procedimiento se llama Credal-C4.5, en donde utiliza una teoría matemática basada en probabilidades imprecisas, y las medidas de incertidumbre. De esta manera, Credal-C4.5 calcula las probabilidades de las características y la variable de clase mediante el uso de probabilidades imprecisas. Además se utiliza un nuevo criterio de división, el cual es llamado Gain Ratio(razon de la ganancia) para generar ganancia de información imprecisa. La aplicación de medidas de incertidumbre sobre conjuntos convexos de distribuciones de probabilidad son llamados (conjuntos Credal). De esta manera, Credal-C4.5 construye árboles para la solución de problemas de clasificación suponiendo que el conjunto de entrenamiento no es totalmente fiable. También se llevaron a cabo varios estudios experimentales que comparan este nuevo procedimiento con otros y se obtiene la siguiente conclusión principal: en los dominios de la clase ruido, Credal-C4.5 obtiene árboles más pequeños y mejor rendimiento que C4.5 clásico.

2.1.1.7. C5.0

C5.0 incorpora nuevas mejoras, tales como la ponderación de distintos casos y tipos de errores de clasificación. En C4.5, todos los errores son tratados como iguales, pero en las aplicaciones prácticas de algunos errores de clasificación son más graves que otros. C5.0 permite un costo independiente que se define para cada par de clases, y luego construye clasificadores para minimizar los costes de clasificación errónea esperados en lugar de las tasas de error.

Los mismos casos también pueden ser de importancia desigual. En una aplicación que clasifica los individuos como probable o improbable, por ejemplo, la importancia de cada caso pueda variar con el tamaño de la cuenta. C5.0 tiene disposición para un atributo de peso caso que cuantifica la importancia de cada caso. Si esto aparece, C5.0 intenta minimizar la tasa de error de predicción ponderada.

C5.0 tiene varios tipos de datos nuevos, además de los disponibles en C4.5, incluyendo fechas, horas, marcas de tiempo, atributos discretos, y las etiquetas de caso. Además de los valores perdidos, C5.0 permite valores que se observan como no aplicable. Además, C5.0 proporciona facilidades para la definición de nuevos atributos como funciones de otros atributos.

Algunas aplicaciones de minería de datos recientes se caracterizan por una muy alta dimensionalidad, con cientos o incluso miles de atributos. C5.0 puede aventar automáticamente los atributos antes de construir un clasificador, descartando aquellas que parecen ser sólo marginalmente relevante. Para aplicaciones de alta dimensión, se puede conducir a clasificadores más pequeños y con una mayor exactitud de predicción, y con frecuencia puede reducir el tiempo requerido para generar conjuntos de reglas (Rulequest, 2010).

Según (Pang, 2009), utilizando el algoritmo C5.0 e incrustando tecnología boosting en la matriz de costos y el árbol de coste razonable se establece un nuevo modelo de evaluación crediticia individual Comercial del Banco.

(Chiang, 2011), propone el algoritmo C5.0 para analizar los datos de implantes para generar reglas de clasificación que pueden ser utilizado como un método de predicción antes de la cirugía de implante basado en los indicadores de rendimiento de precisión, sensibilidad, especificidad.

(Pandya, 2015), sugiere el algoritmo C5.0 como la base clasificadora de modo que el sistema propuesto clasifique el conjunto de resultados con alta precisión y bajo uso de memoria. El proceso de clasificación genera menos reglas, al ser comparado con otras técnicas por lo que el sistema propuesto tiene uso de memoria baja. La tasa de error es baja cuando la precisión en un conjunto de resultados es alta y el árbol podado se genera, por lo que el sistema genera resultados rápidos como comparar con otra técnica. La técnica de selección de funciones supone que los datos contienen muchas características redundantes, de modo que retire características que no proporciona ninguna información útil en cualquier contexto.

(Niu, 2009), utiliza el algoritmo C5.0 para construir un conjunto de clasificadores basados en las características que diferencian a OLTP y OLAP y a continuación, utilizar el clasificador para identificar el tipo de carga de trabajo en el sistema de gestión de base de datos de sintonización (DBMS).

(Bujlow, 2012), sugiere el algoritmo C5.0 para mejorar el rendimiento de la red para superar los inconvenientes en la clasificación de tráfico proveniente de la infraestructura de internet de alta velocidad. La mayor parte del tráfico es generado por los navegadores web, en los diferentes tipos de servicios basados en el protocolo HTTP. El algoritmo C5.0 distingue los diferentes tipos de

aplicaciones con gran precisión formando reglas de clasificación para diferentes tipos de tráfico (Bujlow, Classification of HTTP traffic based on C5.0 Machine Learning Algorithm, 2012). (Pashaei, 2015), menciona que las ventajas de C5.0 cuando lo compararon con otro clasificador árbol de decisión como C4.5 y CART es que C5.0 tiene tasas de error notablemente inferiores. Por lo tanto es más precisa y mucho más rápido y es altamente optimizado.

El algoritmo C5.0 descubre patrones en los datos y los utiliza para hacer predicciones exactas. Clasifica objetos de datos basados en la ganancia de información de sus atributos. A pesar de que responde a datos ruidosos y desaparecidos, su exactitud puede ser mejorada. Por lo cual se propone un puesto de poda algoritmo de árbol de decisión que utilizará C5.0 como su base y la teoría Bayesiana posterior como un potenciador. La poda se realiza mediante la evaluación del árbol de decisión utilizando la teoría Bayesiana posterior. La teoría Bayesiana utiliza la probabilidad para juzgar la validez relativa de hipótesis en cuanto a datos ruidosos e inciertos. El algoritmo propuesto intenta apoyar el uso de memoria baja, una mayor precisión y mejoras en la velocidad con la ayuda de los árboles de decisión más pequeños. También reducirá los riesgos asociados con el exceso de montaje (Mehta, 2015).

2.1.1.8. Classification Regression Trees CART

El término árbol de regresión y clasificación (CART), es un término genérico utilizado para referirse tanto de los procedimientos anteriores, como en los primero introducidos por Breiman. Los árboles utilizados para la regresión y los árboles utilizados para la clasificación tienen algunas similitudes pero también algunas diferencias, tales como el procedimiento utilizado para determinar dónde se dividió (Breiman, 1984).

CART es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos grupos hijos o nodos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado. Las divisiones se seleccionan de modo que “la impureza” de los hijos sea menor que la del grupo madre y estas están definidas por un valor de una variable explicativa (Deconinck, 2006). La regresión no-paramétrica es un procedimiento que requiere un número mínimo de supuestos, donde el ajuste es realizado únicamente a partir de los datos; por esta razón el modelo ajustado por regresión no-paramétrica podría considerarse como la verdadera curva de los datos (Cleveland, 1979).

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos (Timofeev, 2004):

1. Construcción del árbol máximo.
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada (“cross-validation”).

Dentro de las principales ventajas de CART, (Domínguez, 2012) destacan:

1. No necesita hipótesis acerca de la distribución de las variables.
2. Puede trabajar con datos de distintos tipos: categóricos y continuos.
3. Sus resultados son robustos a los outliers.
4. Son invariantes a transformaciones monótonas de los datos, tales como el logaritmo neperiano.
5. Permite combinaciones lineales entre las variables.

6. Selecciona automáticamente las variables que más reducen los errores de clasificación.

No obstante, el mismo (Lewis, 2000) menciona que el CART presenta algunos inconvenientes.

Entre sus principales desventajas se tienen:

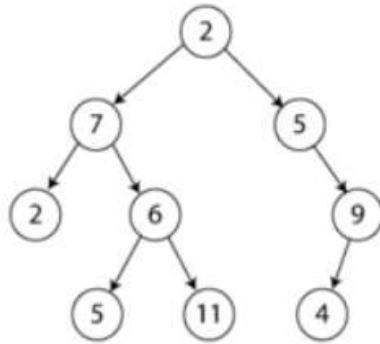


Figura 2. Ejemplo de árbol binario

asociados con las predicciones derivadas usando el algoritmo CART para clasificar a un conjunto de datos.

- Al ser un algoritmo binario, tiende a generar árboles de muchos niveles. Por ello, el árbol resultante puede que no presente los resultados de manera eficiente, sobre todo si la misma variable ha sido utilizada para la división de varios niveles consecutivos.
- CART no está basado en un modelo probabilístico. No existen intervalos de confianza

En el caso particular del algoritmo CART, los árboles obtenidos al ejecutarlo son árboles binarios, es decir, cada nodo del árbol tiene, a lo sumo, dos hijos, como en la Figura 2. Ejemplo de árbol binario.

Una variante de CART es el denominado algoritmo CHAID. La idea subyacente tras este algoritmo es la misma que la que se emplea en el algoritmo CART, con la diferencia de que, en este caso, el árbol de clasificación no tiene por qué ser necesariamente binario sino que puede ser n-ario, es decir, cada nodo puede tener un número de hijos mayor que dos (Lara Torralbo, 2008).

2.1.1.9. Principal Component Analysis PCA

Es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Intuitivamente la técnica sirve para hallar las causas de la variabilidad de un conjunto de datos y ordenarlas por importancia.

Técnicamente, el PCA busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. El PCA se emplea sobre todo en análisis exploratorio de datos y para construir modelos predictivos. El PCA comporta el cálculo de la descomposición en autovalores de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo (Smith, 2002).

El modelo de análisis por componentes principales, es uno de los métodos multivariados de análisis el cual ha sido utilizado con grandes data sets multidimensionales. El uso del análisis por componentes principales permite reducir el número de variables de un data sets, mientras que retiene la variación presente en un data sets. Esta reducción se alcanza tomando las variables y encontrando las combinaciones de ésta para producir componentes principales, no correlacionados (Nie, 2009).

El análisis por componentes principales se puede utilizar en diversos aspectos, uno de ellos es la detección de fallas, lo cual es demandante debido a la creciente tendencia de mejoramiento de los procesos, con el fin de mejorar aspectos de productividad sin olvidar aspectos de la calidad en los procesos, de tal forma que permitan monitorear y predecir posibles fallas que puedan ocurrir durante el proceso. El método de análisis por componentes principales mostró un buen desempeño en la tarea de detección de fallas, en tanto que se pudo comparar con otros métodos, observándose

resultados considerables (Xiaogang, 2008). Otro ejemplo en el que se utiliza para la detección de fallas en los procesos tenemos (Jinyu Guo, 2010).

Otras de las aplicaciones que se le dan al método de análisis por componentes principales, es la del proceso de verificación de firmas, en donde es comparado con otras técnicas, y en el que método de análisis por componentes, muestra buenos resultados (Shih Yin Ooi, 2016).

El Análisis de Componentes Principales (ACP) pertenece a un grupo de técnicas estadísticas multivariantes, eminentemente descriptivas. El enfoque francés de este análisis fue desarrollado por (Benzécri, 1980). Posteriormente ha sido muy difundido, especialmente en el tratamiento de grandes masas de datos. En el trabajo se aplicó el ACP según el criterio de la Escuela Francesa. El ACP permite reducir la dimensionalidad de los datos, transformando el conjunto de p variables originales en otro conjunto de q variables in-correlacionadas ($q < p$) llamadas componentes principales. Las p variables son medidas sobre cada uno de los n individuos, obteniéndose una matriz de datos de orden np ($p < n$).

En el ACP existe la opción de usar la matriz de correlaciones o bien, la matriz de covarianzas. En la primera opción se le está dando la misma importancia a todas y a cada una de las variables; esto puede ser conveniente cuando el investigador considera que todas las variables son igualmente relevantes. La segunda opción se puede utilizar cuando todas las variables tengan las mismas unidades de medida y además, cuando el investigador juzga conveniente destacar cada una de las variables en función de su grado de variabilidad. Las nuevas variables (componentes principales) son obtenidas como combinaciones lineales de las variables originales. Los componentes se ordenan en función del porcentaje de varianza explicada. En este sentido, el primer componente será el más importante por ser el que explica mayor porcentaje de la varianza de los datos. Queda a criterio del investigador decidir cuántos componentes se elegirán en el estudio.

El análisis se realiza en el espacio de las variables y, en forma dual, en el espacio de los individuos. Se acostumbra a representar gráficamente los puntos-variables y los puntos-individuos tomando como ejes de coordenadas los componentes. A veces, puede facilitar la interpretación de los resultados, el observar la similar ubicación de los puntos en los planos respectivos. Aunque el plano de puntos-variables no se superpone al plano de puntos-individuos, es de gran utilidad "interpretar" la cercanía de un grupo de puntos-individuos, a ciertas variables (Foguet, 1988).

En la práctica es frecuente que se disponga de información adicional que amplía la matriz de datos originales. Se puede tener otras medidas de los individuos de la muestra, o también nuevos individuos para los que se conozcan las variables analizadas. A estos datos adicionales se les llama suplementarios o ilustrativos porque no intervienen en la formación de los componentes. En estos casos se calculan las coordenadas de cada punto individuo o variable respecto a los ejes y se representan en los gráficos. Esto permite analizar las relaciones de la información suplementaria con los componentes principales. Generalmente la introducción de estos datos en el análisis se hace porque facilita la interpretación de los resultados (González Martín, 2002).

2.1.1.10. Principal Coordinate Analysis PCoA

PCoA es adecuado para el manejo de una amplia gama de datos, la información relativa a las variables originales no se pueden recuperar. Esto se debe a que PCoA toma una matriz de (des) similitud derivada de los datos originales como entrada y no las propias variables originales. Sin embargo, las puntuaciones de objeto a lo largo del PCoA pueden correlacionarse con puntuaciones de los objetos a lo largo del eje de cada variable original, asumiendo que son o variables cuantitativas o ficticias (Legendre P, 1998). Esto puede ser utilizado como una medida de la contribución de las variables originales a un eje PCoA dado.

El análisis coordinado principal, comienza mediante la proyección de las distancias en el espacio euclidiano en un mayor número de dimensiones. Esto no es difícil; siempre y cuando las distancias sean bastante largas, su comportamiento es bueno, entonces sólo se necesitan $n-1$ dimensiones para con el punto n de datos. PCoA comienza poniendo el primer punto en el origen, y el segundo a lo largo del primer eje la distancia correcta desde el primer punto, a continuación, añade la tercera modo que la distancia a la primera 2 es correcta: esto generalmente significa la adición de un segundo eje . Esto continúa hasta que todos los puntos se añadidos (Gower, 2005). En PCoA, cuando valores propios negativos están presentes en los resultados de descomposición, la matriz de distancia D puede ser modificado utilizando el Lingoés o el procedimiento de Cailliez para producir resultados sin valores propios negativos.

En el procedimiento de (Lingoés, 1971), un c_1 constante, igual a dos veces el valor absoluto del valor negativo más grande de la principal original de coordenadas de análisis, se añade a cada cuadrado de la distancia original en la matriz de distancia, a excepción de los valores diagonales. Un director coordina un nuevo análisis, realizando en las distancias modificadas, tiene como máximo $(n-2)$ valores propios positivos, al menos 2 valores propios nulos y sin valor propio negativo.

En el procedimiento de (Cailliez, 1983), se añade una constante C_2 a las distancias originales en la matriz de distancia, excepto los valores diagonales. El cálculo de C_2 se describe en (Legendre & Gower, 1986). Un nuevo director de coordinar el análisis, realizado en las distancias modificadas, tiene como máximo $(n-2)$ valores propios positivos, al menos 2 valores propios nulos y sin valor propio negativo.

2.1.1.11. Non Metric Multidimensional Scaling

El objeto de no métrica MDS, así como de la métrica MDS, es encontrar las coordenadas de los puntos en p-espacio dimensional, por lo que hay un buen acuerdo entre las proximidades observadas y las distancias entre puntos. El desarrollo de MDS no métrico fue motivada por dos debilidades principales en el MDS métrica (Fahrmeir, 1984):

1. La definición de una conexión funcional explícita entre diferencias y distancias con el fin de derivar las distancias de disimilitudes dado, y
2. La restricción de la geometría euclidiana con el fin de determinar las configuraciones de objetos.

Escalamiento multidimensional no métrico (MDS, también NMDS y SNM) es una técnica de ordenación que difiere en varios aspectos de casi todos los otros métodos de coordinación. En la mayoría de ordenación métodos, muchos ejes se calculan, pero sólo unos pocos son vistos, debido a las limitaciones gráficas.

En MDS, un pequeño número de ejes se eligen de manera explícita antes del análisis y los datos están instalados en esas dimensiones; no hay ejes ocultos de variación. En segundo lugar, la mayoría de otros métodos ordenación son analíticos y por lo tanto resultan en una sola solución única a un conjunto de datos. En contraste, MDS es una técnica numérica que busca iterativamente una solución y paradas cómputo cuando se ha encontrado una solución aceptable, o se detiene después de un determinado pre- Número de intentos. Como resultado, una ordenación MDS no es una solución única y una subsiguiente Análisis MDS en el mismo conjunto de datos y siguiendo la misma metodología probablemente resultar en una ordenación algo diferente.

En tercer lugar, MDS no es un valor propio-vector propio técnica como análisis de componentes principales o análisis de correspondencias que coordina la los datos de tal manera que el eje 1 explica la mayor cantidad de varianza, el eje 2 se explica el siguiente mayor cantidad de varianza, y así sucesivamente. Como resultado, una ordenación MDS se puede girar, invertido, o centrada a cualquier configuración deseada (Holand, 2008).

Las aplicaciones del NMDS incluyen la visualización científica y la minería de datos en campos como la ciencia cognitiva, ciencias de la información, la psicofísica, la psicometría, la comercialización y la ecología. Las nuevas aplicaciones se plantean en el ámbito de nodos inalámbricos autónomos que pueblan un espacio o un área. MDS puede aplicar como un enfoque mejorado en tiempo real para el seguimiento y la gestión de tales poblaciones.

Por otra parte, el MDS se ha utilizado ampliamente en la geo-estadística, para modelar la variabilidad espacial de los patrones de una imagen (representándolos como puntos en un espacio de menor dimensión), (Cambria, 2013) y el procesamiento del lenguaje natural, para modelar la relación semántica y afectiva de los conceptos de lenguaje natural (mediante la representación de ellos como puntos en un espacio vectorial 100 - dimensional) (Honarkhah, 2010).

MDS se está convirtiendo en un método popular usado en la secuencia de la agrupación y visualización. En la bioinformática, MDS se usa para reducir la dimensionalidad dando las puntuaciones de disimilitud de cada par de secuencias. Estas puntuaciones de disimilitud se calculan generalmente utilizando Secuencia de alineación. Mediante la cartografía de cada secuencia del espacio de alta dimensión a un espacio visualmente aceptable (por ejemplo, el espacio 2D / 3D), las correlaciones entre cada grupo de secuencia se pueden observar fácilmente (Yang, 2012).

En la mercadeo, MDS es una técnica estadística para la toma de las preferencias y percepciones de los encuestados y los representa en una rejilla visual, llamada mapas perceptuales. Mediante la

cartografía de múltiples atributos y múltiples marcas, al mismo tiempo, un mayor conocimiento del mercado y de las percepciones de los consumidores se puede lograr, en comparación con una de dos básica atribuyen mapa perceptual (Fripp, 2014).

2.1.2. Distribuciones Estadísticas

La distribución de probabilidad calcula el conjunto de valores de la variable aleatoria, y la probabilidad de que el valor de la variable aleatoria esté dentro de un rango definido (Rivera, 2002).

2.1.2.1. Bernoulli

Las características de un experimento aleatorio Bernoulli son:

1. El experimento tiene solamente dos posibles resultados mutuamente excluyente denominados éxito (E) y fracaso (F). De esta manera el espacio muestral es dado por

$$S = \{ \text{éxito, fracaso} \} \quad (5)$$

2. La probabilidad de éxito y fracaso son constantes y se denotan por p y q ($q=1-p$) respectivamente

2.1.2.2. Variable aleatoria Bernoulli y su función de probabilidad

Una variable aleatoria Bernoulli X se define como el resultado numérico de una prueba Bernoulli o de manera formal como una función

$$\begin{aligned} X: S &\rightarrow R \\ \text{exito} &\rightarrow 1 \\ \text{fracaso} &\rightarrow 0 \end{aligned} \quad (6)$$

y así el rango de la variable aleatoria es $\{1,0\}$, el cual es denotado como $X=\{1,0\}$.

Una variable aleatoria de Bernoulli, por sí sola, tiene poco interés en las aplicaciones de ingeniería. En cambio la realización de una serie de experimentos Bernoulli conduce a varias distribuciones de probabilidad discretas muy útiles.

La función de probabilidad de una variable Bernoulli es dada por

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & x = 0,1 \\ 0 & \text{en otro caso} \end{cases} \quad (7)$$

Donde

p Es la probabilidad de éxito en una sola prueba.

x Es el número de éxitos en la prueba.

El **parámetro** es $p, 0 \leq p \leq 1$.

2.1.2.3. Media y Varianza de la Distribución Bernoulli

La media y varianza de una variable aleatoria Bernoulli son respectivamente

$$E(X) = p \quad \text{y} \quad V(X) = p(1-p) \quad (8)$$

Ejemplo:

El experimento de seleccionar un producto y observar si tiene defectos o no. Aquí se puede definir ser defectuoso como el éxito y no ser defectuoso como el fracaso.

2.1.2.4. Distribución Binomial

Es una de las distribuciones de probabilidad más útiles (control de calidad, producción, investigación). Tiene que ver con el experimento aleatorio que produce en cada ensayo o prueba uno de dos resultados posibles mutuamente excluyentes: ocurrencia de un criterio o característica específico (llamado éxito) y no ocurrencia de éste (llamado fracaso). Los términos o calificativos de "éxito y fracaso" son solo etiquetas y su interpretación puede no corresponder con el resultado positivo o negativo de un experimento en la realidad.

2.1.2.5. Criterios o Propiedades de la Distribución Binomial

1. El experimento aleatorio consiste en n ensayos o pruebas repetidas, e idénticas y fijadas antes del experimento (pruebas de Bernoulli). Son pruebas con reemplazamiento o con reposición.
2. Cada uno de los n ensayos o pruebas arroja solo uno de dos resultados posibles resultados: éxito o fracaso.
3. La probabilidad del llamado éxito (ocurrencia)= P , permanece constante para cada ensayo o prueba.
4. Cada prueba o ensayo se repite en idénticas condiciones y es independiente de las demás.
5. El interés recae en hallar la probabilidad de obtener x número de éxitos al realizar n ensayos del mismo E.A.

Cuando estas propiedades se cumplen en el experimento aleatorio se dice que el constituye un proceso de Bernoulli y cada uno de los ensayos que lo conforman se llama experimento de Bernoulli.

La función de probabilidad de X en esas condiciones será:

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{para } x = 0, 1, 2, \dots, n \\ 0 & \text{en otro caso} \end{cases} \quad (10)$$

Para n entero y $0 \leq p \leq 1$

2.1.2.6. Distribución Multinomial

$$(X_1, X_2, \dots, X_k) \sim M(n, p_1, p_2, \dots, p_k) \quad (11)$$

Definición:

Es una distribución de probabilidad conjunta para múltiples variables aleatorias (X_1, X_2, \dots, X_k) discretas donde cada $X_i \sim b(n, p_i)$, dándose cuando en cada prueba ó ensayo independiente (con reposición) del E.A. interesa contar el número de éxitos en cada una de la k maneras como se puede dar un atributo.

Explicación

El atributo calidad de un producto se puede dar como: Excelente, bueno, regular y malo.

1. Son n pruebas o ensayos repetidos e idénticos (con reposición).
2. En cada prueba o ensayo se pueden producir k resultados.
3. Las probabilidades de cada uno de los k resultados (p_1, p_2, \dots, p_k) permanecen constantes en todas las pruebas o ensayos.
4. Son pruebas o ensayos independientes.
5. El interés se centra en contar los X_1, X_2, \dots, X_k éxitos que se producen en los n ensayos de cada una de las k categorías posibles de observar cada vez.

Si una prueba o intento puede dar cualquiera de los k resultados posibles E_1, E_2, \dots, E_k con probabilidades p_1, p_2, \dots, p_k , entonces la distribución Multinomial dará la probabilidad de que:

$$\left. \begin{array}{l} E_1 \text{ ocurra } x_1 \text{ veces: } P[X_1 = x_1] \\ E_2 \text{ ocurra } x_2 \text{ veces: } P[X_2 = x_2] \\ \cdot \\ \cdot \\ \cdot \\ E_k \text{ ocurra } x_k \text{ veces: } P[X_k = x_k] \end{array} \right\} \quad (12)$$

En n pruebas independientes, y donde: $x_1 + x_2 + \dots + x_k = n$ y

$$p_1 + p_2 + \dots + p_k = 1.0. \quad (13)$$

Como son pruebas independientes, cualquier orden específico que produzca

$$\left\{ \begin{array}{l} x_1 \text{ resultados para } E_1 \\ x_2 \text{ resultados para } E_2 \\ \cdot \\ \cdot \\ \cdot \\ x_k \text{ resultados para } E_k \end{array} \right\} \quad (14)$$

Ocurrirá con $p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$ de probabilidad. El número de órdenes o arreglos que pueden producir

resultados similares será:
$$\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1! x_2! \dots x_k!} \quad (15)$$

Combinando los dos componentes, se tiene entonces que:

$$f(x_1, x_2, \dots, x_k) = P[X_1 = x_1, X_2 = x_2, \dots, X_k = x_k,] = \binom{n}{x_1, x_2, \dots, x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

Con $\sum_{i=1}^k x_i = n$ y $\sum_{i=1}^k p_i = 1.0.$ (16)

2.1.2.7. Distribución de Poisson

$$XP(\lambda)$$

Llamada así por su autor Simeón Denis Poisson, probabilista del siglo XIX, pues fue el primero en describirla. Es una generalización de la distribución binomial cuando sobre un $E. A.$ Se define una variable aleatoria X que representa el número de éxitos independientes que ocurren para intervalos de medida específicos (tiempos, lugares, espacios), además con una probabilidad de ocurrencia pequeña. Se le llama distribución de los "eventos raros" pues se usa como aproximación a la binomial cuando el tamaño de muestra es grande y la proporción de éxitos es pequeña. Esos intervalos de medida pueden referirse a: Tiempo: (Segundo, minuto, hora, día, semana, etc.) Área: (Segmento de línea, pulgada cuadrada, Centímetro cuadrado, etc.). Volumen:(Litro, galón, onza, etc.)

Ejemplo

- Número de defectos por m^2 en piezas similares de un material.
- Número de personas que llegan a un taller automotriz en un lapso de tiempo específico.
- Número de impulsos electrónicos errados transmitidos durante espacio de tiempo específico.
- Número de llamadas telefónicas que ingresan a un conmutador por minuto.
- Número de interrupciones en servicios de energía en intervalos de un día.
- Cantidad de átomos que se desintegran en sustancia radioactiva.
- Número de accidentes automovilísticos en un cruce específico durante una semana.

2.1.2.8. Criterios o Propiedades de la Distribución de Poisson

- Se da un intervalo de medida que divide un todo de números reales y donde el conteo de ocurrencias es aleatorio. Esa división puede ser un subintervalo de medida.
- El número de ocurrencias o de resultados en el intervalo o subintervalo de medida, es independiente de los demás intervalos o subintervalos. por eso se dice que el proceso de Poisson no tiene memoria.
- La probabilidad de que un solo resultado ocurra en un intervalo de medida muy corto ó pequeño es la misma para todos los demás intervalos de igual tamaño y es proporcional a la longitud del mismo o al tamaño de medida.
- La probabilidad de que más de un resultado ocurra en un intervalo o subintervalo corto es tan pequeña que se considera insignificante (cerca o igual a cero).

Procesos que se ajustan a estos criterios, se dice, son procesos de Poisson.

Definición

Sea una variable aleatoria que representa el número de eventos aleatorios independientes que X ocurren con igual rapidez en un intervalo de medida. Se tiene entonces que la función de probabilidad de esta variable, se expresa por:

$$f(x) = P[X = x] = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x = 0, 1, 2, \dots; \\ 0 & \text{en cualquier otro punto ó valor} \end{cases} \quad (17)$$

Donde λ es parámetro de tendencia central de la distribución y representa el número promedio o cantidad esperada de ocurrencias (éxitos) del evento aleatorio por unidad de medida o por muestra;

$e = 2.71828$ y $x =$ Número de ocurrencias específicas para el cual se desea conocer la

probabilidad respectiva. Según sea el valor de $\lambda > 0$, se define toda una familia de probabilidades

de Poisson. La probabilidad de que una variable aleatoria de Poisson X sea menor ó igual a un valor de x se halla por la función de distribución acumulativa, planteada entonces como:

$$p(X \leq x) = F(x) = \sum_{k=0}^x \frac{e^{-\lambda} \lambda^k}{k!} \quad (18)$$

Los resultados de las probabilidades individuales para valores de X serán más pequeños conforme la variable aleatoria toma valores cada vez más grandes.

2.1.2.9. Distribución Normal

La distribución normal es de suma importancia en estadística por tres razones principales:

- Numerosas variables continuas de fenómenos aleatorios tienden a comportarse probabilísticamente mediante ésta.
- Es el límite al que convergen tanto variables aleatorias continuas como discretas.
- Proporciona la base de la inferencia estadística clásica debido a su relación con el teorema del límite central.

Propiedades de la distribución normal

- Su grafica tiene forma acampanada.
- El valor esperado, la mediana y la moda tienen el mismo valor cuando la variable aleatoria se distribuye normalmente.
- Su *dispersión media* es igual a 1.33 desviaciones estándar. Es decir, el alcance inter-cuartil está contenido dentro de un intervalo de dos tercios de una desviación estándar por debajo de la media a dos tercios de una desviación estándar por encima de la media.

En la práctica, algunas de las variables que observamos sólo pueden aproximar estas propiedades. Así que si el fenómeno puede medirse aproximadamente mediante la distribución normal se tendrá:

- Que el polígono puede verse en forma de campana y simétrico.
- Sus mediciones de tendencia central tienen bastante parecido.
- El valor inter-cuartil puede diferir ligeramente de 1.33 desviaciones estándar.
- El dominio de la variable aleatoria normalmente distribuida generalmente caerá dentro de 3 desviaciones estándar por encima y por debajo de la media.

El modelo matemático

$$\text{Notación: } N(\mu, \sigma^2), X \sim N(\mu, \sigma^2) \quad (19)$$

El modelo o expresión matemática que representa una función de densidad de probabilidad se denota mediante el símbolo $f(x)$. Para la distribución normal, se tiene la siguiente función de probabilidad.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (20)$$

Donde

e Es la constante matemática aproximada por 2.71828

π Es la constante matemática aproximada por 3.14159

Parámetros $\left\{ \begin{array}{l} \mu_X \text{ es el valor esperado de la variable aleatoria} \\ \sigma_X \text{ es la desviación estándar de la variable aleatoria} \end{array} \right.$

x Es cualquier valor de la variable aleatoria continua, donde $-\infty < x < +\infty$

Así,

$$E(X) = \mu_X \quad (21)$$

$$Var(X) = \sigma_X^2 \quad (22)$$

2.1.2.10. Distribución F

La distribución F es una distribución de probabilidad continua. También se le conoce como distribución F de Snedecor (por George Snedecor) o como distribución F de Fisher-Snedecor (por Ronald Fisher).

Una variable aleatoria de distribución F se construye como el siguiente cociente:

$$F = \frac{U_1/d_1}{U_2/d_2} \quad (23)$$

Donde

- U1 y U2 siguen una distribución chi-cuadrado con d1 y d2 grados de libertad respectivamente, y
- U1 y U2 son estadísticamente independientes.

La distribución F aparece frecuentemente como la distribución nula de una prueba estadística, especialmente en el análisis de varianza. Véase el test F.

La función de densidad de una F (d1, d2) viene dada por

$$g(x) = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1 x}{d_1 x + d_2} \right)^{d_1/2} \left(1 - \frac{d_1 x}{d_1 x + d_2} \right)^{d_2/2} x^{-1} \quad (24)$$

Para todo número real $x \geq 0$, donde d1 y d2 son enteros positivos, y B es la función beta.

La función de distribución es

$$G(x) = I_{\frac{d_1 x}{d_1 x + d_2}}(d_1/2, d_2/2) \quad (25)$$

2.1.3. Análisis De La Varianza Con Un Factor (Anova)

El análisis de la varianza permite contrastar la hipótesis nula de que las medias de K poblaciones ($K > 2$) son iguales, frente a la hipótesis alternativa de que por lo menos una de las poblaciones difiere de las demás en cuanto a su valor esperado (Riera, 2005). Este contraste es fundamental en el análisis de resultados experimentales, en los que interesa comparar los resultados de K 'tratamientos' o 'factores' con respecto a la variable dependiente o de interés.

$$\begin{aligned} H_0: \mu_1 = \mu_2 = \dots = \mu_K = \mu \\ H_1: \exists \mu_j \neq \mu \quad j = 1, 2, \dots, K \end{aligned} \quad (26)$$

El Anova requiere el cumplimiento los siguientes supuestos:

- Las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son normales.
- Las K muestras sobre las que se aplican los tratamientos son independientes.
- Las poblaciones tienen todas igual varianza (homocedasticidad).
- El ANOVA se basa en la descomposición de la variación total de los datos con respecto a la media global (SCT), que bajo el supuesto de que H_0 es cierta es una estimación de σ^2 obtenida a partir de toda la información muestral, en dos partes:
- Variación dentro de las muestras (SCD) o Intra-grupos, cuantifica la dispersión de los valores de cada muestra con respecto a sus correspondientes medias.
- Variación entre muestras (SCE) o Inter-grupos, cuantifica la dispersión de las medias de las muestras con respecto a la media global.

Las expresiones para el cálculo de los elementos que intervienen en el Anova son las siguientes:

$$\text{Media Global: } \bar{X} = \frac{\sum_{j=1}^K \sum_{i=1}^{n_j} x_{ij}}{n} \quad (27)$$

$$\text{Variación Total: } \text{SCT} = \sum_{j=1}^K \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2 \quad (28)$$

Variación Intra-grupos:
$$SCD = \sum_{j=1}^K \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2 \quad (29)$$

Variación Inter-grupos:
$$SCE = \sum_{j=1}^K (\bar{X}_j - \bar{X})^2 n_j \quad (30)$$

Siendo x_{ij} el i -ésimo valor de la muestra j -ésima; n_j el tamaño de dicha muestra y \bar{X}_j su media.

Cuando la hipótesis nula es cierta $SCE/K-1$ y $SCD/n-K$ son dos estimadores insesgados de la varianza poblacional y el cociente entre ambos se distribuye según una F de Snedecor con $K-1$ grados de libertad en el numerador y $N-K$ grados de libertad en el denominador. Por lo tanto, si H_0 es cierta es de esperar que el cociente entre ambas estimaciones será aproximadamente igual a 1, de forma que se rechazará H_0 si dicho cociente difiere significativamente de 1.

2.1.4. Permanova

Es un algoritmo que mide la respuesta simultánea de una o varias variables a uno o varios factores en un diseño experimental ANOVA, basado en cualquier medida de distancias y usando análisis de permutaciones.

- Está basado en cualquier medida de distancias y usando análisis de permutaciones.
- Es una técnica de análisis multivariante (sobre medidas de distancia) con varios factores (balanceados) (MANOVA).
- Se le aplica análisis de permutaciones sobre las matrices de distancia (PERmutaciones).

2.1.5. Pruebas No Paramétricas

Se denominan pruebas no paramétricas aquellas que no presuponen una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre (*distribution free*). En la mayor parte de ellas los resultados estadísticos se derivan únicamente a partir de procedimientos de ordenación y recuento, por lo que su base lógica es de fácil comprensión. Cuando trabajamos con muestras pequeñas ($n \leq 10$) en las que se desconoce si es válido suponer la

normalidad de los datos, conviene utilizar pruebas no paramétricas, al menos para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal. En estos casos se emplea como parámetro de centralización la **mediana**, que es aquel punto para el que el valor de X está el 50% de las veces por debajo y el 50% por encima.

2.1.5.1. Prueba de Wilcoxon de los Rangos con Signo

Esta prueba nos permite comparar nuestros datos con una mediana teórica (por ejemplo un valor publicado en un artículo).

Llamemos M_0 a la mediana frente a la que vamos a contrastar nuestros datos, y sea $X_1, X_2 \dots X_n$ los valores observados. Se calcula las diferencias $X_1 - M_0, X_2 - M_0, \dots, X_n - M_0$. Si la hipótesis nula fuera cierta estas diferencias se distribuirían de forma simétrica en torno a cero.

Para efectuar esta prueba se calculan las diferencias en valor absoluto $|X_i - M_0|$ y se ordenan de menor a mayor, asignándoles su rango (número de orden). Si hubiera dos o más diferencias con igual valor (empates), se les asigna el rango medio (es decir que si tenemos un empate en las posiciones 2 y 3 se les asigna el valor 2.5 a ambas). Ahora calculamos R_+ la suma de todos los rangos de las diferencias positivas, aquellas en las que X_i es mayor que M_0 y R_- la suma de todos los rangos correspondientes a las diferencias negativas. Si la hipótesis nula es cierta, ambos estadísticos deberán ser parecidos, mientras que si nuestros datos tienen a ser más altos que la mediana M_0 , se reflejará en un valor mayor de R_+ , y al contrario si son más bajos. Se trata de contrastar si la menor de las sumas de rangos es excesivamente pequeña para ser atribuida al azar, o, lo que es equivalente, si la mayor de las dos sumas de rangos es excesivamente grande.

2.1.5.2. Prueba de Wilcoxon para Contrastar Datos Pareados

El mismo razonamiento lo podemos aplicar cuando tenemos una muestra de parejas de valores, por ejemplo antes y después del tratamiento, que podemos denominar $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. De la misma forma, ahora calcularemos las diferencias $X_1 - Y_1, X_2 - Y_2, \dots, X_n - Y_n$ y las ordenaremos en valor absoluto, asignándoles el rango correspondiente. Calculamos R_+ la suma de rangos positivos (cuando X_i es mayor que Y_i), y la suma de rangos negativos R_- . Ahora la hipótesis nula es que esas diferencias proceden de una distribución simétrica en torno a cero y si fuera cierta los valores de R_+ y R_- serán parecidos.

2.1.5.3. Prueba de Mann-Whitney para Muestras Independientes

Si tenemos dos series de valores de una variable continua obtenidas en dos muestras independientes: $X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m$, procederemos a ordenar conjuntamente todos los valores en sentido creciente, asignándoles su rango, corrigiendo con el rango medio los empates. Calculamos luego la suma de rangos para las observaciones de la primera muestra S_x , y la suma de rangos de la segunda muestra S_y . Si los valores de la población de la que se extrajo la muestra aleatoria de X se localizan por debajo de los valores de Y , entonces la muestra de X tendrá probablemente rangos más bajos, lo que se reflejará en un valor menor de S_x del teóricamente probable. Si la menor de las sumas de rangos es excesivamente baja, muy improbable en el caso de que fuera cierta la hipótesis nula, ésta será rechazada.

Existen más pruebas no paramétricas de entre las que a continuación mencionamos las más habituales, remitiendo al lector interesado a cualquier libro básico de bioestadística:

- Prueba de Kruskal-Wallis para comparar K muestras
- Prueba de Friedman para comparar K muestras pareadas (bloques)
- Coeficiente de correlación de Spearman para rangos
- Prueba de rachas de Wald-Wolfowitz

2.1.6. Pruebas T

La prueba estadística t de Student para muestras dependientes es una extensión de la utilizada para muestras independientes. De esta manera, los requisitos que deben satisfacerse son los mismos, excepto la independencia de las muestras; es decir, en esta prueba estadística se exige dependencia entre ambas, en las que hay dos momentos uno antes y otro después. Con ello se da a entender que en el primer período, las observaciones servirán de control o testigo, para conocer los cambios que se susciten después de aplicar una variable experimental.

Con la prueba t se comparan las medias y las desviaciones estándar de grupo de datos y se determina si entre esos parámetros las diferencias son estadísticamente significativas o si sólo son diferencias aleatorias (Raymundo, 2015).

Consideraciones para su uso

- El nivel de medición, en su uso debe ser de intervalo o posterior.
- El diseño debe ser relacionado.
- Se deben cumplir las premisas para métricas.

En cuanto a la homogeneidad de varianzas, es un requisito que también debe satisfacerse y una manera práctica es demostrarlo mediante la aplicación de la prueba ji cuadrada de Bartlett. Este procedimiento se define por medio de la siguiente fórmula:

$$t = \frac{\bar{d}}{\frac{\sigma d}{\sqrt{N}}}$$

Dónde:

t = valor estadístico del procedimiento. \bar{d} = Valor promedio o media aritmética de las diferencias entre los momentos antes y después. sd = desviación estándar de las diferencias entre los momentos antes y después. N = tamaño de la muestra.

La media aritmética de las diferencias se obtiene de la manera siguiente:

$$\bar{d} = \frac{\sum d}{N} \quad (31)$$

La desviación estándar de las diferencias se logra como sigue:

$$\sigma d = \sqrt{\frac{\sum (d - \bar{d})^2}{N - 1}} \quad (32)$$

Pasos:

1. Ordenar los datos en función de los momentos antes y después, y obtener las diferencias entre ambos.
2. Calcular la media aritmética de las diferencias (\bar{d}).
3. Calcular la desviación estándar de las diferencias (sd).
4. Calcular el valor de t por medio de la ecuación.
5. Calcular los grados de libertad (gl) $gl = N - 1$.
6. Comparar el valor de t calculado con respecto a grados de libertad en la tabla respectiva, a fin de obtener la probabilidad.
7. Decidir si se acepta o rechaza la hipótesis.

2.2. Trastornos Hipertensivos en las Mujeres Embarazadas (Normales)

Preeclampsia

Se considera preeclampsia cuando hay presión arterial sistólica (PAS) ≥ 140 y/o presión arterial diastólica (PAD) ≥ 90 mmHg en 2 tomas aparte de 6 horas de intervalo asociado de proteinuria ≥ 300 mg en 24 horas.

Eclampsia

Cuando estas pacientes presentan convulsiones tónico-clónicas generalizadas, se denomina eclampsia.

Hipertensión gestacional:

Aparición de hipertensión (PA sistólica ≥ 140 mmHg y/o PA diastólica ≥ 90 mmHg) sola sin proteinuria después de 20 semanas de gestación.

Hipertensión crónica:

Hipertensión (PA sistólica ≥ 140 mmHg y/o PA diastólica ≥ 90 mmHg) diagnosticada antes del embarazo o en la primera mitad del embarazo (< 20 semanas de gestación) y que continúe durante > 12 semanas después del parto.

Tipos de Preeclampsia

La preeclampsia se clasifica en severa y no severa según las cifras de tensión arterial (Tabla1). Se denomina preeclampsia severa, si presenta presión arterial sistólica ≥ 160 mmHg o presión arterial diastólica ≥ 110 mmHg (Figura 3.)

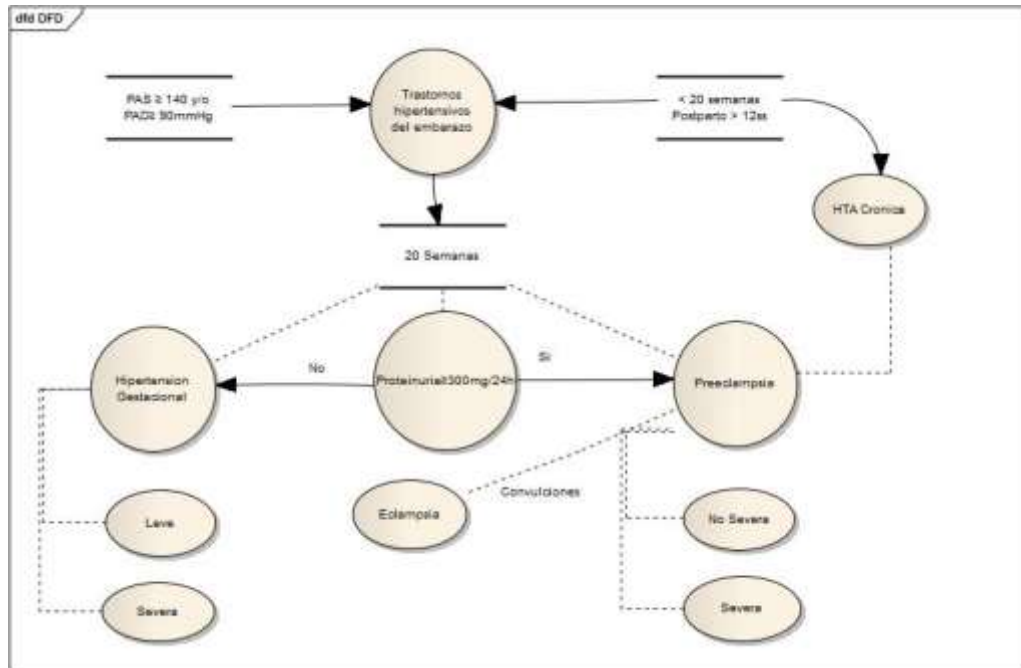


Figura 3. Algoritmo diagnóstico para la clasificación de Trastornos hipertensivos

Tomado de Texto de Obstetricia y Ginecología (Fecolsog, 2010)

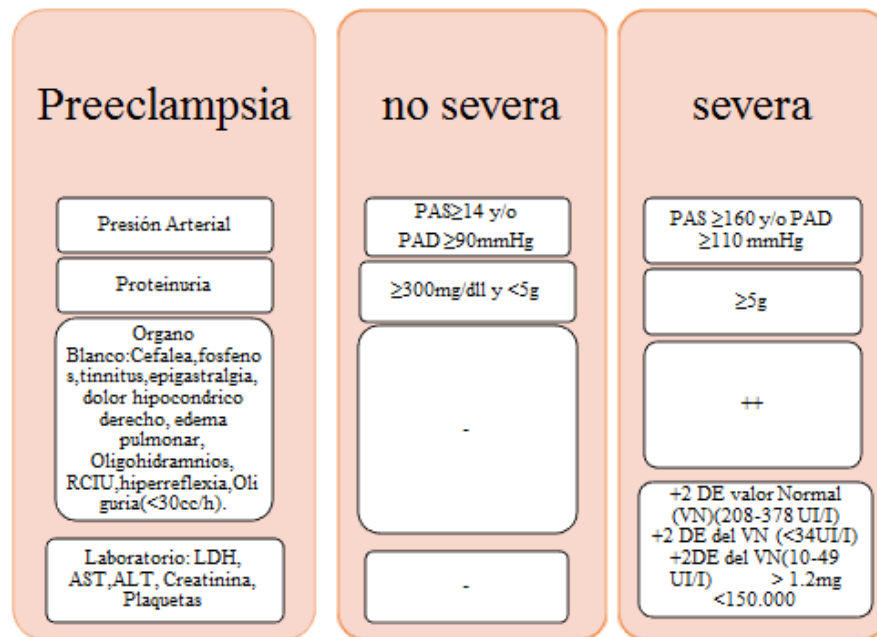


Figura 4. Clasificación de Preeclampsia

Tomado de Texto de Obstetricia y Ginecología (Fecolsog, 2010)

PE No Severa

Aparición de hipertensión (PA sistólica ≥ 140 mmHg y/o PA diastólica ≥ 90 mmHg) y proteinuria después de la semana 20 de gestación

PE Severa

PE más uno de los siguientes criterios:

- PA sistólica ≥ 160 mmHg y/o PA diastólica ≥ 110 mmHg (en dos ocasiones en un intervalo de ≥ 6 horas).
- Proteinuria (> 5 g proteína /24 horas o tira reactiva $\geq 3+$ en dos muestras de orina recogidas con un intervalo mínimo de 4 horas).
- Deterioro de la función renal (creatinina sérica $\geq 1,2$ mg/dl salvo que se sepa que anteriormente era elevada u oliguria < 500 ml/24 horas).
- Edema pulmonar.
- Deterioro de la función hepática.
- Síntomas neurológicos (molestias cerebrales o visuales, cefalea severa).
- Trastornos hematológicos (trombocitopenia, hemólisis).
- CIR.

Preeclampsia Precoz y Tardía:

- PE precoz: aparición de la enfermedad $< 34+0$ semanas de gestación.
- PE tardía: aparición de la enfermedad $\geq 34+0$ semanas de gestación.

2.3. Factores a Diagnosticar en la preeclampsia

En la práctica clínica actual, no existe un método óptimo para seleccionar a aquellas gestantes con un mayor riesgo de desarrollar PE. Las gestantes con factores conocidos de muy alto riesgo son

seguidas de forma más intensiva en consultas especializadas, según se recomiendan los protocolos de la Federación colombiana de Asociaciones de Obstetricia y Ginecología.

Los factores a diagnosticar en el embarazo en el estudio fueron los siguientes:

- **Hipertensión:** PA sistólica ≥ 140 mmHg y/o PA diastólica ≥ 90 mmHg (en dos Ocasiones con un intervalo ≥ 6 horas) .
- **Proteinuria:** análisis de proteínas en orina de 24 horas ($\geq 0,3$ g proteínas/ 24 horas), en caso de emergencia si no se puede determinar proteínas en orina de 24 horas, la determinación de proteínas se realizará en una muestra puntual de orina (≥ 30 mg/dl de proteínas) o ratio proteína/ creatinina (≥ 30 mg proteína/mmol creatinina). En caso de urgencia se utilizará una tira reactiva ($\geq 2+$).
- **LDH:** La lactato deshidrogenasa (o también llamada "deshidrogenasa del ácido láctico" (LDH)). Se mide con mayor frecuencia para verificar daño tisular y debe estar entre los valores normales de 105 a 333 UI/L (unidades internacionales por litro).
- **AST:** La aspartato aminotransferasa. Se mide con mayor frecuencia para verificar el nivel de enzimas en la sangre y para estar en valores normales su rango debe estar entre 10 y 34 UI/L.
- **Creatinina:** La creatinina es un compuesto orgánico generado a partir de la degradación de la creatina (que es un nutriente útil para los músculos). Se mide el nivel de creatinina en la sangre y se hace para ver qué tan bien funcionan los riñones. Los valores normales deben estar entre 0.6 a 1.1 mg/dL para las mujeres.
- **Plaquetas:** Las plaquetas tienen una función importantísima en la coagulación de la sangre. Su disminución leve de plaquetas apenas tienen síntomas, pero a medida que van disminuyendo pueden aparecer importantes hemorragias nasales (epistaxis

), moratones (púrpura) o ante cualquier herida, dificultad en parar el sangrado. Pero donde realmente tiene importancia es en el momento del parto, ya que tanto el parto por vía vaginal como por cesárea pueden ocasionar importantes hemorragias. Se mide el nivel de plaquetas. Los valores normales deben estar entre 150.000 a 450.000 plaquetas por centímetro cúbico de sangre.

- **Edad gestacional:** La determinación de la edad gestacional clásicamente fue basada en el número de semanas de amenorrea, la cual fija como criterios absolutos ciclos regulares de 28 días, no dudas en la fecha exacta y no uso de anticonceptivos por lo menos tres meses previos. Sin embargo, solo 50% cumple con estos criterios, por lo que la determinación actual de la edad gestacional debe hacerse en base a la ecografía del primer trimestre o ser confirmada con esta última. El momento más exacto y confiable para determinar la edad gestacional es entre las 8 y 12 semanas. La evaluación ecográfica del primer trimestre incluye la medida de LCN, que es el método más exacto para la estimación de la edad gestacional (Redondo, 2011).
- **Peso:** Durante el embarazo es muy importante el control del peso materno para prevenir enfermedades (como la diabetes gestacional) y complicaciones en el parto (por feto muy grande).

Los intervalos de incremento de peso recomendados (para feto único) son:

- bajo peso materno: 12 –18 Kg.
- peso materno normal: 11 –15 Kg.
- sobrepeso materno: 6 – 7 Kg.
- obesidad materna: < 6 Kg.

- **Método de Concepción:**

Método en que fue concebido el embarazo (espontáneo, drogas hormonales, inseminación artificial etc).

- **Edad de la paciente:**

La edad materna avanzada incrementa el riesgo de PE. Por otra parte estas pacientes tienen mayor incidencia de factores de riesgo adicionales como diabetes o hipertensión crónica. Un estudio demográfico realizado en EEUU sugiere que el riesgo de PE aumenta un 30% por cada año adicional a partir de los 34 años. Las edades inferiores no han mostrado afectar al riesgo de PE (Alfaro).

2.3.1. Factores de Riesgo

Los factores de riesgo de PE han sido clasificados o divididos de diferente manera por varios autores. Así, (Serrano NC, 2005) y otros⁶ los dividen en genéticos y medioambientales, mientras que (Contreras F, 2003) y otros en preconceptionales o crónicos y vinculados con el embarazo.

- **Maternos:**

Preconceptionales:

- Edad materna menor de 20 y mayor de 35 años
- Étnicos.
- Historia personal de PE (en embarazos anteriores).
Presencia de algunas enfermedades crónicas: hipertensión arterial, obesidad, diabetes mellitus, resistencia a la insulina, enfermedad renal, neurofibromatosis, síndrome antifosfolípido primario (anticuerpos antifosfolípidos) y otras enfermedades autoinmunes (síndrome antifosfolípido secundario), trombofilias y dislipidemia.

- **Relacionados con la gestación en curso:**

- Primigravidez o embarazo de un nuevo compañero sexual.
- Sobredistención uterina (embarazo gemelar y polihidramnios).
- Embarazo molar en nulípara.
- Ambientales:
 - Malnutrición por defecto o por exceso.
 - Escasa ingesta de calcio previa y durante la gestación.
 - Hipomagnesemia y deficiencias de zinc y selenio.
 - Alcoholismo durante el embarazo.
 - Bajo nivel socio económico.
 - Cuidados prenatales deficientes.
 - Estrés crónico.

2.4. Efectos de la Preeclampsia

Las manifestaciones siguientes o *síntomas premonitorios*, en presencia de preeclampsia, sus manifestaciones o factores de riesgo, nos sugieren la posibilidad de que la paciente convulsione en cualquier momento:

- Dolor de cabeza (50-75%.)
- Alteraciones visuales: estrellitas en los ojos, visión borrosa, molestia con la luz (19-32%)
- Dolor en la boca del estómago
- Trastornos mentales y neurológicos

2.5. Tratamiento de la Preeclampsia Grave

- Hospitalización, en una sala oscura y aislada de ruidos
- Reposo absoluto, de preferencia en de cúbito lateral izquierdo
- Régimen normosódico
- Se controlarán los signos vitales cada 4 hs, el peso materno una vez al día, la medición de la diuresis y un movidograma diario.
- Sedación con diazepam (dosis de ataque: 10 mgrs diluido en 10cc de dextrosa 5% EV lento) .
- Sulfato de magnesio: la dosis de ataque es de 4 a 5 grs. en 500 de dextrosa al 5% a goteo libre. La dosis de mantenimiento en de 5 grs. en 500 cc de dextrosa al 5% a 35 gotas por minuto (equivale a razón de 1gr por hora).
- Durante el uso del sulfato de magnesio es necesario mantener: reflejos presentes, diuresis mayor a 25 ml/hora y ausencia de depresión respiratoria.
- Hipotensores por vía parenteral frente a la falta de respuesta a los antihipertensivos orales Y se deberá tener siempre presente la posibilidad de interrupción del embarazo, siendo los criterios para la interrupción del
- mismo los sig.: preeclampsia moderada con feto maduro (edad gestacional mayor a las 37 semanas)
- Preeclampsia severa con edad gestacional mayor de 34 semanas Preeclampsia severa con feto inmaduro, en que fracasa el tratamiento médico o se presenta el deterioro progresivo del estado materno (HTA severa, crisis hipertensiva) .
- Evidencia de deterioro de la unidad feto placentaria, independientemente de la edad gestacional.
- Presencia de eclampsia

2.6. Tratamiento de la Preeclampsia Leve

Se realizará un tratamiento en forma ambulatoria, debe alertarse a la paciente sobre los signos y síntomas de empeoramiento de la preeclampsia. Debe recomendarse una dieta regular, sin restricciones de sal, ni limitaciones en la actividad física. Además debe indicarse la toma de la presión arterial en forma diaria, la vigilancia del peso y los edemas como así también la realización de laboratorio de control en forma periódica (Dra. Verónica Natalia Joerin, 2007). Tratamiento de la preeclampsia moderada:

- Hospitalización
- Reposo, de preferencia en decúbito lateral izquierdo
- Régimen completo, normosódico
- Control de signos vitales maternos y LCF cada 4 horas
- Sedación con diazepam oral (5mg cada 4 horas)
- Medición del peso y la diuresis diaria
- Hipo tensores orales si la presión diastólica es mayor a 100 mmHg. Deberán usarse drogas como hidralazina, alfa metil dopa, labetalol o antagonistas del calcio.
- La dosis recomendada para la alfa metil dopa es de 500-2000 mg/día (entre 250 a 500 mgrs. c/ 6 hs).

Si a pesar de estas medidas no se logra un buen control de las cifras tensionales y aparecen signos de mayor daño materno (elevación de la proteinuria, deterioro del cléarance de creatinina) o fetal, evidenciado a través de los parámetros de evaluación de la unidad feto placentaria, debe plantearse la interrupción del embarazo.

Cuando la evolución del cuadro hipertensivo señala la conveniencia de interrumpir el embarazo, y se trata de gestaciones menores de 34 semanas, con pulmón fetal inmaduro, es conveniente inducir la maduración fetal con corticoides, e interrumpir la gestación a las 48 hs de la primer dosis.

CAPÍTULO 3

3. ASPECTOS METODOLÓGICOS

En este capítulo se presentaron todos los metodológicos y técnicas usadas para la obtención de la información. Además se delimita el objeto de estudio de este proyecto y los procedimientos utilizados para poder llevar a cabo la investigación

3.1. Tipo y Diseño de la Investigación

El proyecto necesitará un método de investigación cualitativo y uno cuantitativo, explicados a continuación:

- Análisis cualitativo de la investigación corresponde a la captura de los datos con mayor relevancia utilizados en los métodos tradicionales para prevención de la preeclampsia. De tal forma que los pacientes firman un documento de consentimiento informado.
- Análisis cuantitativo sobre el modelo de predicción de la preeclampsia durante el segundo trimestre de gestación. Para esto se diseñará una prueba basada en el uso de datos sobre mujeres embarazadas que presentaron preeclampsia y otras que no presentaron para determinar el nivel de eficiencia del método de predicción.

Los pasos metodológicos se describen a continuación.

- Hacer un recorrido del estado del arte sobre técnicas usadas comúnmente para predicción de preeclampsia en el segundo trimestre, y los criterios útiles para

comparar cuáles técnicas son las mejores, como precisión de la predicción, facilidad de implementación, etc.

- Determinar el grupo de variables adecuado para predecir la enfermedad
- Diseñar un algoritmo que basado en una técnica de inteligencia artificial pueda predecir la preeclampsia con niveles más eficientes que las técnicas actualmente usadas
- Implementar un prototipo de software que implemente el algoritmo
- Evaluar la eficiencia del algoritmo en la predicción usando datos reales de mujeres y comparar con métodos normalmente usados.
- Diseñar e implementar un software para asistir al médico y a la mujer en un sano proceso de gestación, que integre la información sobre predicción de la enfermedad.

3.2. Área de Estudio

Comprende el Área Gineco-Obstétrica perteneciente la Clínica Portoazul de la ciudad de Barranquilla entre el mes de Noviembre a Diciembre 2015.

3.3. Universo

El universo del estudio fue seleccionado a través de las historias clínicas de mujeres embarazadas con diagnóstico de Preeclampsia atendidas en la Clínica Portoazul.

3.4. Muestra

La muestra esperada es de 50 casos de mujeres preeclámpticas y 50 casos de mujeres embarazadas sin preeclampsia, considerando los criterios de inclusión y exclusión.

3.5. Criterios de Inclusión

Mujeres embarazadas sin preeclampsia y con preeclampsia leve o severa.

3.6. Criterios de Exclusión

Mujeres eclámpticas, hipertensión crónica, hipertensión transitoria o tardía.

3.7. Instrumento de Recolección de la Información

Para la recolección de los datos de investigación se elaboró un formulario donde se recogió toda la información necesaria de acuerdo a los objetivos planteados; utilizando como fuente los médicos de la Clínica Portoazul, se elaboró el listado de variables de las historias clínicas de las mujeres embarazadas con diagnóstico de Preeclampsia, en el período establecido.

3.8. Descripción de Procedimientos

- Aprobación del proyecto por el comité de Ética de la Universidad del Norte.
- Autorización de datos predictores a partir de las historias clínicas proyecto la clínica Portoazul mediante previa solicitud.
- Estudio sistemático de la información obtenida.
- Análisis de resultados

CAPÍTULO 4

4. MODELO DE PREDICCIÓN DE PREECLAMPSIA

En este capítulo se presenta el diseño del modelo de detección de preeclampsia, basado en clasificación variables predictoras usando los algoritmo C5.0 y NMDS que ayudaran a crear un conjunto de reglas de clasificación, provenientes del entrenamiento con el fin de caracterizar patrones de detección de preeclampsia que ayuden al experto en obstetricia a examinar y verificar más fácilmente los resultados obtenidos para apoyar la toma de decisiones.

El capítulo comienza planteando el problema que se tiene en la detección de preeclampsia, luego se hace la descripción del modelo propuesto de detección, el cual no pretende cubrir todos los aspectos involucrados en un proceso KDD, sino solo aquellos que permitan representar el conocimiento a extraer en términos de reglas de clasificación a partir de un árbol de decisión. De esta manera, el modelo presenta un conjunto de fases que permiten la preparación de datos, extracción, predicción o clasificación y presentación del conocimiento, comunes en cualquier proceso de minería de datos.

4.1. Modelos de Extracción de Conocimiento para la Predicción de Preeclampsia

En este trabajo se plantea un modelo de detección de preeclampsia, basado en análisis de datos utilizando técnicas de minería de datos, específicamente para un conjunto de datos de varias entidades prestadoras de servicios de salud(hospitales). El modelo se presenta en la Figura5, la cual presenta a grandes rasgos la iteración de diferentes módulos y la comunicación con elementos externos al sistema (Archivos). El modelo se compone de varios módulos que proporcionan una

funcionalidad específica, con el fin de obtener conocimiento de la relación que existe entre el conjunto de datos. Los módulos mencionados son:

1. Preparación y selección de datos : Encargado de la recolección de la fuente de información, y preparación del conjunto de datos para la extracción del conocimiento.

2. Extracción del conocimiento. Es el impulsor del modelo de detección propuesto , y esta conformado

por :

- Algoritmo de Aprendizaje. Para el proceso se construye un árbol de decisión a través del algoritmo C5.0 generando un conjunto de reglas de clasificación. El C5.0 divide la muestra en función del campo que ofrece la mayor ganancia de información. Las distintas sub-muestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las sub-muestras de nuevo. Por último se vuelven a examinar las divisiones del nivel inferior, y se eliminan o podan las que no contribuyen significativamente con el valor del modelo.
- Eliminación de factores. Para obtener la mayor ganancia y efectividad dentro del modelo, se eliminan los factores que no aportaron ganancia de información dentro del modelo.

3. Predicción: Encargado de inferir una predicción sobre un conjunto de datos nuevos. Para ello se parte del conocimiento encontrado con el algoritmo C5.0 para hacer la clasificación del conjunto de datos. Luego con la técnica NMDS explicado en el capítulo 5.2 se observa la separación de los datos y posteriormente se comparan los resultados con los obtenidos por el algoritmo C5.0 para la toma de decisión .

4. Presentación de la información final al doctor: se muestran los valores obtenidos de la predicción.

Los pasos se muestran en la Figura 5. Modelo de extracción de conocimiento para la detección de Preeclampsia Los pasos 1,2 se explican en el capítulo 4 mientras que los pasos 3 ,4 se explican en el capítulo 5.

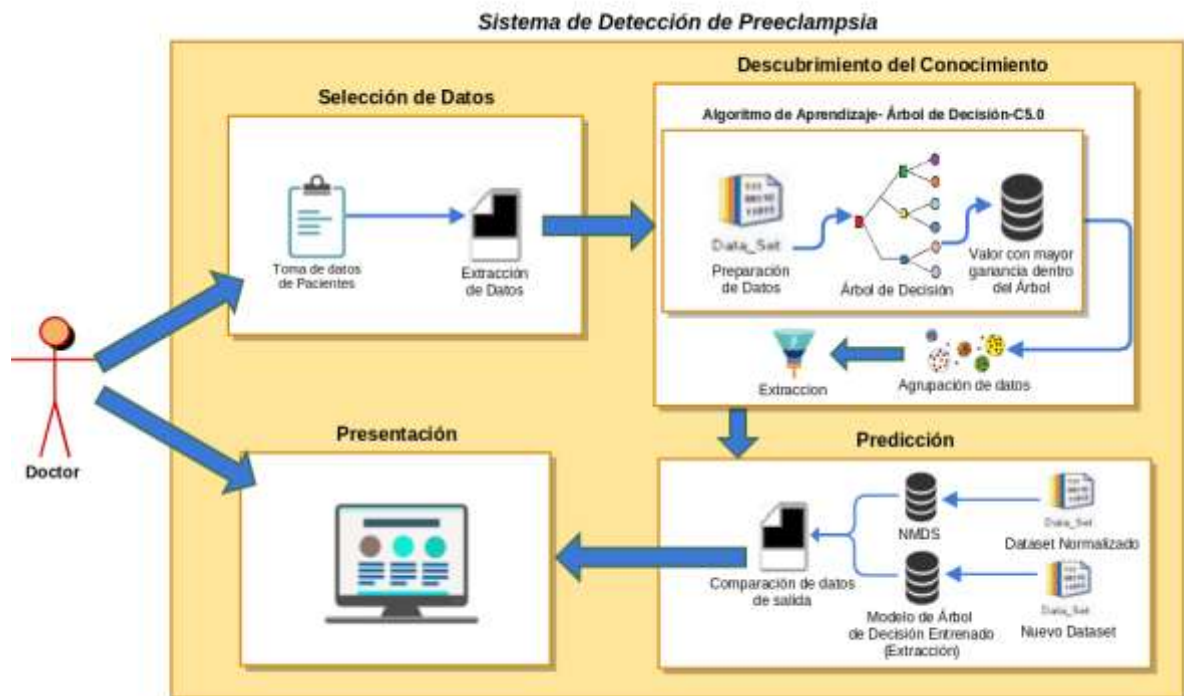


Figura 5. Modelo de extracción de conocimiento para la detección de Preeclampsia

4.2. Preparación y Selección de Datos

En esta fase se determinan, seleccionan y extraen los datos de la fuente de información para el proceso de extracción de conocimiento. En esta etapa se realizan las siguientes actividades.

1. Extracción de datos relevantes de la fuente de información haciendo énfasis en:

- Identificación de los datos relevantes en las fuentes de información (para el caso, un modelo relacional de tablas)

- Extracción e integración. A pesar de que los aspectos importantes a evaluar del modelo pueden estar distribuidos en diferentes tablas en una misma base de datos, en el modelo propuesto, se considera que estos datos conforman una sola tabla, a partir de una consulta en la base de datos que retorna una tabla virtual (conjunto de registros), con los datos a utilizar en el proceso de extracción de conocimiento.

2. Preparación de datos, en esta parte se hace énfasis en:

- Eliminación de registros inconsistentes.
- Eliminación de atributos que no aportan información, por ejemplo redundancia de información, valores perdidos, etc.
- Ruido e inconsistencia de datos.
- Discretización de atributos.

3. Selección de la información a utilizar en el proceso de extracción de conocimiento, a partir de determinar cuáles van hacer las variables independientes y cuyavariablen será objetivo (conclusión). Para esto, el usuario es el encargado de determinar las variables a considerar en el proceso de aprendizaje y extracción de reglas.

4. Presentación de los resultados del aprendizaje

Esta fase puede ser soportada por herramientas estadísticas, herramientas de limpieza y otras, de tal manera que el usuario pueda realizar estas operaciones de una forma eficiente.

4.3. Algoritmo de Extracción

A continuación se va a explicar el algoritmo C5.0.

Ejemplos de aprendizaje.

Atributo_salida: Atributo a predecir por el árbol.

Atributos: Lista de atributos a comprobar por el árbol.

(1) Crear una raíz para el árbol.

(2) Si todos los ejemplos son positivos Retornar (raíz, + (positiva))

(3) Si todos los ejemplos son negativos Retornar (raíz, - (negativa))

(4) Si Atributos = \emptyset Retornar (raíz, l)

(Donde l es el máximo valor común de Atributo_salida en Ejemplos) Si ninguna de las anteriores condiciones se cumple

Inicio

(1) Seleccionar el atributo A con mayor Ganancia (Ejemplos,A)

(2) El atributo de decisión para raíz es A

(3) Para cada posible valor V_i de A

(3.1) Añadir una rama a raíz con el test $A=V_i$

(3.2) Ejemplos_ V_i es el subconjunto de Ejemplos con valor V_i

para A

(3.3) Si Ejemplos_ V_i = \emptyset

Entonces añadir un nodo (n,l) a partir de la rama creada.

(l es el máximo valor común de Atributo_salida

en Ejemplos).

Sino añadir a la rama creada el subárbol

C5.0(Ejemplos_ v_i , Atributo_salida, Atributos-{A})

Fin

Retorna (raíz)

CAPÍTULO 5.

5. IMPLEMENTACIÓN DEL MODELO

En este capítulo se presenta la implementación del modelo de detección de preeclampsia, basado en clasificación variables predictoras usando los algoritmo C5.0 y NMDS y en el cual prodemos observar la estructura de la información ingresada, aplicabilidad de los algoritmos y resultados y graficas de resultados .

5.1. Algoritmo C5.0 en R

- 1) Se leen los datos y se almacena en la variable `datas` y se imprime

```
require(gdata)  
library(gdata)  
read.xls(/home/topor/Descargas/prueba.xls)  
datas= read.xls(/home/topor/Descargas/prueba.xls)
```

	edad	pam	Proteinuria	Idh	ast	creatinina	plaquetas	drogas	eg	clase
1	45	117.0	411	449	45	0.63	151	si	37.0	preeclampsia
2	24	101.0	580	215	15	0.53	233	no	31.2	preeclampsia
3	20	100.0	241	198	16	0.60	308	no	34.4	preeclampsia
4	21	60.0	143	232	123	0.53	252	no	33.0	preeclampsia
5	32	137.0	406	408	35	0.71	190	no	31.0	preeclampsia
6	44	109.0	491	300	39	0.72	376	no	34.0	preeclampsia
7	27	116.0	144	325	30	1.50	295	no	35.0	preeclampsia
8	42	118.0	242	355	51	0.71	130	no	35.0	preeclampsia
9	31	101.0	227	293	49	0.82	352	no	36.0	preeclampsia
10	45	86.0	141	276	18	1.37	368	no	35.0	preeclampsia
11	31	124.0	482	329	53	1.64	123	no	33.0	preeclampsia
12	31	94.0	493	341	18	1.35	383	no	30.0	preeclampsia
13	39	134.0	377	233	43	0.60	148	no	36.0	preeclampsia
14	36	120.0	232	276	32	1.21	187	no	37.0	preeclampsia
15	19	108.0	281	246	15	1.43	274	no	35.0	preeclampsia
16	23	101.0	133	405	31	0.69	318	no	33.0	preeclampsia
17	33	98.0	466	251	25	0.51	366	no	36.0	preeclampsia
18	37	114.0	543	233	42	1.54	232	no	32.0	preeclampsia
19	39	75.0	494	440	52	1.70	350	no	32.0	preeclampsia
20	22	88.0	369	277	45	1.59	120	no	30.0	preeclampsia
21	28	68.0	318	367	43	1.10	139	no	30.0	preeclampsia
22	40	94.0	125	227	35	1.73	354	no	36.0	preeclampsia

Figura 6. Dataset Preeclampsia

- 2) Vemos los primeros datos de la variable del dataset como se observa en la Figura7.

`head(datas)`

```
  edad pam Proteinuria ldh ast creatinina plaquetas drogas  eg   clase
1   45 117          411 449 45         0.63      151    si 37.0 preclampsia
2   24 101          580 215 15         0.53      233    no 31.2 preclampsia
3   20 100          241 198 16         0.60      308    no 34.4 preclampsia
4   21  60          143   9 123         0.53      252    no 33.0 preclampsia
5   32 137          406 408 35         0.71      190    no 31.0 preclampsia
6   44 109          491 300 39         0.72      376    no 34.0 preclampsia
> |
```

Figura 7. Encabezado del Dataset

- 3) Vemos la estructura de la información almacenada como se observa en la Figura8.

`str(datas)`

```
'data.frame':  100 obs. of  10 variables:
 $ edad      : int  45 24 20 21 32 44 27 42 31 45 ...
 $ pam       : num  117 101 100 60 137 109 116 118 101 86 ...
 $ Proteinuria: int  411 580 241 143 406 491 144 242 227 141 ...
 $ ldh       : int  449 215 198 9 408 300 325 355 293 276 ...
 $ ast       : int  45 15 16 123 35 39 30 51 49 18 ...
 $ creatinina : num  0.63 0.53 0.6 0.53 0.71 0.72 1.5 0.71 0.82 1.37 ...
 $ plaquetas  : int  151 233 308 252 190 376 295 130 352 368 ...
 $ drogas     : Factor w/ 2 levels "no","si": 2 1 1 1 1 1 1 1 1 1 ...
 $ eg        : num  37 31.2 34.4 33 31 34 35 35 36 35 ...
 $ clase     : Factor w/ 2 levels "preclampsia",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Figura 8. Estructura del Dataset

- 4) Se utilizan los factores de clasificación cruzada para construir una tabla de recuento de la variable clase en cada combinación de los niveles de los factores

`table(datas$class)`

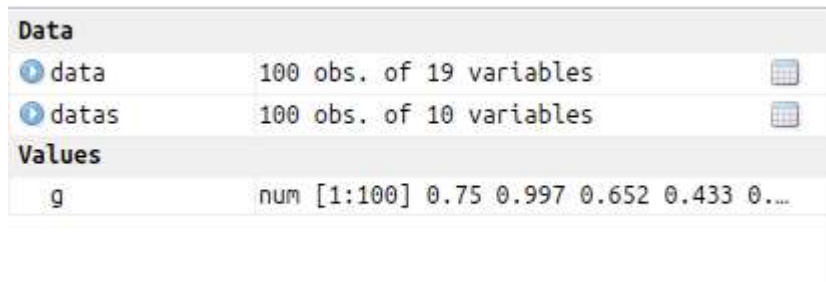
```
preclampsia sinpreclampsia
          50             50
```

- 5) Generamos números aleatorios para asegurar todos los posibles resultados.

```
set.seed(9850)
```

- 6) Se generan valores de un distribución uniforme a partir de los factores

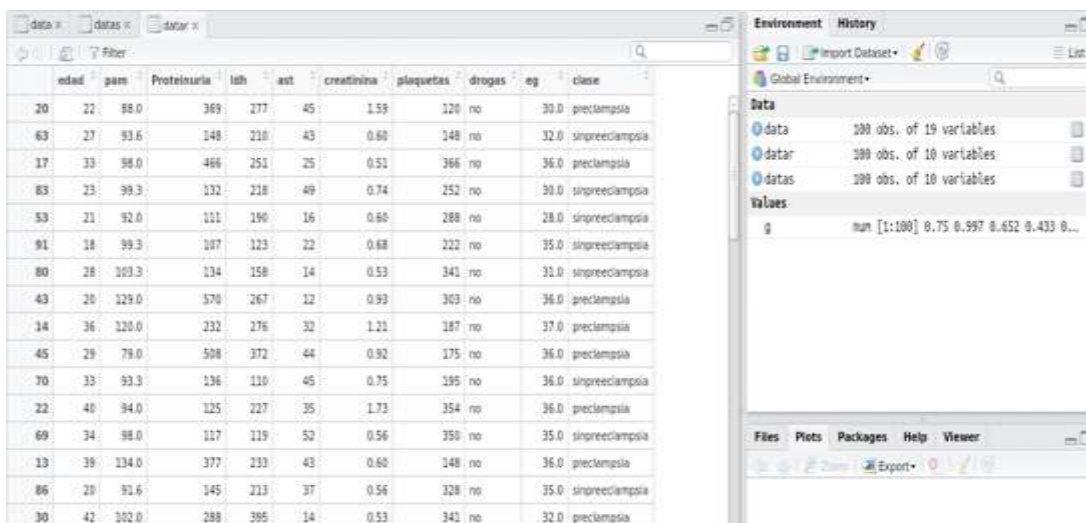
```
g <- runif(nrow(datas))
```



The screenshot shows the R Studio Environment pane. Under the 'Data' section, there are two objects: 'data' with 100 observations of 19 variables, and 'datas' with 100 observations of 10 variables. Under the 'Values' section, the variable 'g' is shown as a numeric vector of length 100, with the first few values being 0.75, 0.997, 0.652, 0.433, and 0.000.

- 7) Se ordena la trama de datos a partir de los valores de la distribución generados

```
datar <- datas[order(g),]
```



The screenshot shows the R Studio interface. The main window displays a data table with 19 columns: edad, pam, Proteinsuria, lsh, ast, creatinina, plaquetas, drogas, eg, and clase. The table contains 19 rows of data. The Environment pane on the right shows three data objects: 'data' (100 obs. of 19 variables), 'datar' (100 obs. of 19 variables), and 'datas' (100 obs. of 10 variables). The 'Values' section shows the variable 'g' as a numeric vector of length 100, with the first few values being 0.75, 0.997, 0.652, 0.433, and 0.000.

- 8) Se realiza la clasificación en 80 registros de datasets a partir de los datos ordenados de la distribución para cada factor

```
m1 <- C5.0(datar[1:80,-10], datar[1:80,10])
```

```
m1      List of 16
names : chr "Generated using R version 3.2.3 (201...
cost : chr ""
costMatrix : NULL
caseWeights : logi FALSE
control : List of 11
..$ subset : logi TRUE
..$ bands : num 0
..$ winnow : logi FALSE
..$ noGlobalPruning : logi FALSE
..$ CF : num 0.25
..$ minCases : num 2
..$ fuzzyThreshold : logi FALSE
..$ sample : num 0
..$ earlyStopping : logi TRUE
..$ label : chr "outcome"
..$ seed : int 690
trials : Named num [1:2] 1 1
.. attr(*, "names")= chr [1:2] "Requested" "Actual"
rbn : logi FALSE
boostResults : NULL
size : int 2
dims : int [1:2] 80 9
call : language C5.0.default(x = datar[1:80, -10], ...
levels : chr [1:2] "preclampsia" "sinpreclampsia"
output : chr "\nC5.0 [Release 2.07 GPL Edition] \tW...
tree : chr "id=\See5/C5.0 2.07 GPL Edition 2010-02...
predictors : chr [1:9] "edad" "pam" "Proteinuria" "...
rules : chr ""
attr(*, "class")= chr "C5.0"
```

- 9) Se imprime el resultado del entrenamiento del algoritmo c5.0

```
m1
```

```
Call:
C5.0.default(x = datar[1:80, -10], y = datar[1:80, 10])

Classification Tree
Number of samples: 80
Number of predictors: 9

Tree size: 2

Non-standard options: attempt to group attributes
```

- 10) Se produce el resumen de los resultados de la clasificación, mostrando el nivel del árbol generado en la clasificación

summary(m1)

```
Console -> C5.0

Call:
C5.0.default(x = datar[1:80, -10], y = datar[1:80, 10])

C5.0 [Release 2.07 GPL Edition] Wed Feb 10 13:08:21 2016
-----

Class specified by attribute 'outcome'

Read 80 cases (10 attributes) from undefined.data

Decision tree:

ldh <= 218: sinpreeclampsia (44/2)
ldh > 218: preclampsia (36)

Evaluation on training data (80 cases):

Decision Tree
-----
Size      Errors
  2      2( 2.5%) <<

(a) (b) <-classified as
---- ----
 36   2 (a): class preclampsia
    42 (b): class sinpreeclampsia

Attribute usage:
100.00% ldh

Time: 0.0 secs
```

11) Se predice la clase a través del árbol generado por el algoritmo C5.0

```
p1<-predict(m1, datar[80:100,])
p1
```

```
[1] sinpreeclampsia sinpreeclampsia sinpreeclampsia preclampsia preclampsia preclampsia preclampsia
[8] sinpreeclampsia preclampsia sinpreeclampsia preclampsia sinpreeclampsia preclampsia sinpreeclampsia
[15] preclampsia preclampsia sinpreeclampsia preclampsia preclampsia sinpreeclampsia sinpreeclampsia
Levels: preclampsia sinpreeclampsia
```

12) Construye una tabla de recuento de la variable clase en cada combinación de los niveles de los factores de p1

```
table(datar[80:100,10], Predicted= p1)
```

	Predicted	
	preclampsia	sinpreeclampsia
preclampsia	11	1
sinpreeclampsia	0	9

13) Se imprime el árbol generado en la clasificación como se observa en la Figura9.

```
plot(m1)
```

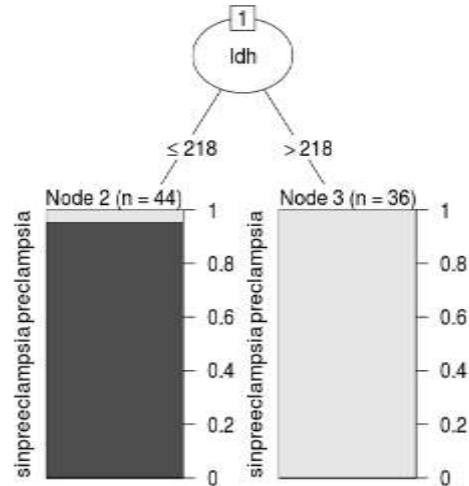


Figura 9. Arbol entrenado C5.0

5.2. Sistema de predicción NMDS en R

5.2.1. Principal Component Analysis PCA

1) Se leen los datos con la función `read.xls` y se le asigna a un objeto llamado `data`

```
library(gdata)  
read.xls("/home/topor/Descargas/ppca.xls")  
data=read.xls("/home/topor/Descargas/ppca.xls")
```

2) Se mira la estructura de `data`

```
str(data)
```

3) Se eliminan las filas con NA

```
data=data[1:8]  
str(data)
```

```

'data.frame': 100 obs. of 8 variables:
 $ edad      : int  45 24 20 21 32 44 27 42 31 45 ...
 $ pam       : num  117 101 100 60 137 109 116 118 101 86 ...
 $ Proteinuria: int  411 580 241 143 406 491 144 242 227 141 ...
 $ ldh       : int  449 215 198 9 408 300 325 355 293 276 ...
 $ ast       : int  45 15 16 123 35 39 30 51 49 18 ...
 $ creatinina : num  0.63 0.53 0.6 0.53 0.71 0.72 1.5 0.71 0.82 1.37 ...
 $ plaquetas  : int  151 233 308 252 190 376 295 130 352 368 ...
 $ eg        : num  37 31.2 34.4 33 31 34 35 35 36 35 ...

```

4) Se mira la correlación entre variables con la función cor() como se observa en la Figura 10.

```
cor(data[,1:8])
```

```

          edad      pam Proteinuria      ldh      ast
edad      1.00000000  0.08490485  0.35086216  0.4416447 -0.001078331
pam       0.08490485  1.00000000  0.28643601  0.3222493 -0.198112048
Proteinuria 0.35086216  0.28643601  1.00000000  0.5962791 -0.086765348
ldh       0.44164471  0.32224934  0.59627912  1.00000000 -0.126749731
ast      -0.001078331 -0.19811205 -0.08676535 -0.1267497  1.000000000
creatinina 0.332785012  0.11904879  0.56724138  0.4210237 -0.052287927
plaquetas -0.048160307 -0.23446893 -0.08358099 -0.1000539 -0.133090816
eg        0.125165161  0.23243593  0.12881324  0.1494961 -0.047834023
 creatinina  plaquetas      eg
edad      0.33278501 -0.04816031  0.12516516
pam       0.11904879 -0.23446893  0.23243593
Proteinuria 0.56724138 -0.08358099  0.12881324
ldh       0.42102375 -0.10005394  0.14949608
ast      -0.05228793 -0.13309082 -0.04783402
creatinina 1.00000000 -0.09937160  0.09040401
plaquetas -0.09937160  1.00000000  0.11114020
eg        0.09040401  0.11114020  1.00000000

```

Figura 10. Correlación entre variables

5) Se ajusta un PCA con la función prcomp() como se observa en la Figura 11.

```
pca1=prcomp(data[,1:8])
```

```
pca1
```

```
Standard deviations:
[1] 176.3690227 84.6914189 72.1734634 17.6355777 11.3601806 6.7683297
[7] 2.3997643 0.3075987
```

```
Rotation:
          PC1          PC2          PC3          PC4
edad    -0.017236187  0.0025455650  2.349211e-02 -0.0351491231
pam     -0.023500073  0.0324717516  1.832285e-02  0.2380928704
Proteinuria -0.913495224 -0.1127120827 -3.907208e-01 -0.0043194852
ldh     -0.401450791  0.1117295657  9.077617e-01 -0.0301178651
ast     -0.009945955  0.0290904487 -2.639906e-02 -0.9698174407
creatinina -0.001253191  0.0001295767 -2.909451e-05 -0.0003166462
plaquetas 0.058394195 -0.9863533244  1.473494e-01 -0.0237954309
eg      -0.002005931 -0.0035276686  3.112888e-03  0.0061959237
          PC5          PC6          PC7          PC8
edad    -0.059174305  0.996611832 -0.0336781322 -0.0061754175
pam     0.966725590  0.063096231 -0.0531337506  0.0025874009
Proteinuria -0.009502447 -0.007032687 -0.0002487547 -0.0011256786
ldh     -0.021321478 -0.030984492 -0.0008256525 -0.0003677505
ast     -0.239595902 -0.019433394 -0.0052845121  0.0005159091
creatinina -0.002819130  0.006127586  0.0038837052  0.9999688663
plaquetas 0.037227351  0.001238972 -0.0056258357  0.0003169556
eg      0.050942264  0.036844077  0.9979815555 -0.0039581460
```

Figura 11. Dataset en Funcion del PCA

6) Se muestra la cabecera de la variable pca1

`str(pca1)`

```
List of 5
 $ sdev   : num [1:8] 176.4 84.7 72.2 17.6 11.4 ...
 $ rotation: num [1:8, 1:8] -0.01724 -0.0235 -0.9135 -0.40145 0.00995 ...
 ..- attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:8] "edad" "pam" "Proteinuria" "ldh" ...
 .. ..$ : chr [1:8] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:8] 29.1 100 251.3 237.5 34 ...
 ..- attr(*, "names")= chr [1:8] "edad" "pam" "Proteinuria" "ldh" ...
 $ scale   : logi FALSE
 $ x       : num [1:100, 1:8] -238.4 -293.6 27.4 191.5 -215.5 ...
 ..- attr(*, "dimnames")=List of 2
 .. ..$ : NULL
 .. ..$ : chr [1:8] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
```

7) Se imprime el resumen del modelo

`summary(pca1)`

```
Importance of components:
          PC1          PC2          PC3          PC4          PC5          PC6
Standard deviation 176.3690 84.6914 72.1735 17.63558 11.36018 6.76833
Proportion of Variance 0.7073 0.1631 0.1184 0.00707 0.00293 0.00104
Cumulative Proportion 0.7073 0.8704 0.9888 0.99589 0.99883 0.99987
          PC7          PC8
Standard deviation 2.39976 0.3076
Proportion of Variance 0.00013 0.0000
Cumulative Proportion 1.00000 1.00000
```

8) Se dibujan los dos primeros factores del PCA como se observa en la Figura12.

`biplot(pca1)`

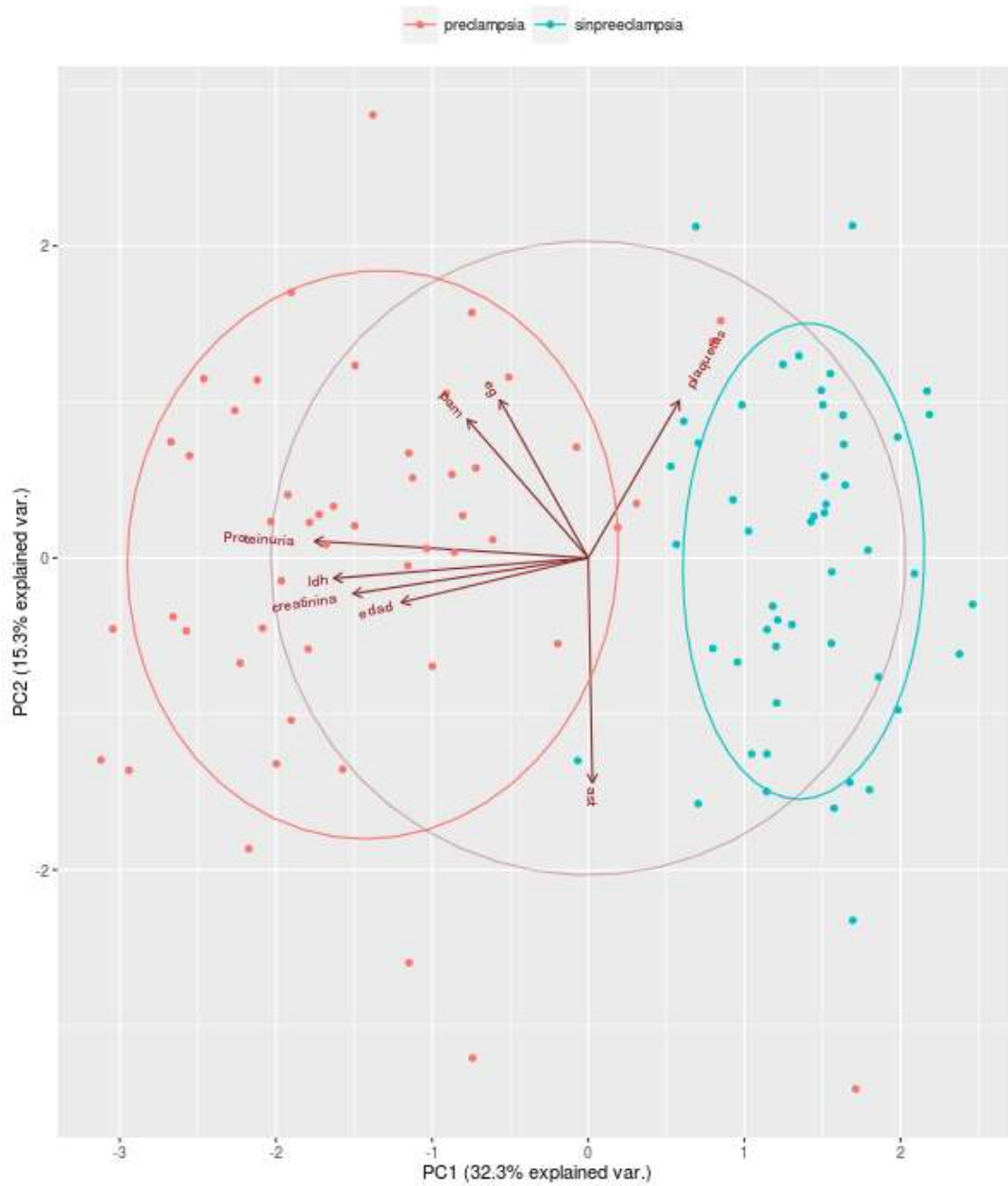


Figura 12. Separación Factores predictores por clase

Es posible notar de la anterior figura la separación de los factores, dentro de los cuales, los componentes principales (ldh, proteinuria, edad y creatinina) se alejan en el mismo sentido de las demás variables del modelo de datos.

5.2.2. Principal Coordinate Analysis PCoA o Non Metric Dimensional Scaling

- 1) Se utiliza la librería vegan y normalizan los datos

```
library(vegan)
```

```
dta=dat[1:8]
```

```
dta
```

	edad	pam	Proteinuria	ldh	ast	creatinina	plaquetas	eg	clase
1	45	117.0	411	449	45	0.63	151	37.0	preclampsia
2	24	101.0	580	215	15	0.53	233	31.2	preclampsia
3	20	100.0	241	198	16	0.60	308	34.4	preclampsia
4	21	60.0	143	233	123	0.53	252	33.0	preclampsia
5	32	137.0	406	408	35	0.71	190	31.0	preclampsia
6	44	109.0	491	300	39	0.72	376	34.0	preclampsia
7	27	116.0	144	325	30	1.50	295	35.0	preclampsia
8	42	118.0	242	355	51	0.71	130	35.0	preclampsia
9	31	101.0	227	293	49	0.82	352	36.0	preclampsia
10	45	86.0	141	276	18	1.37	368	35.0	preclampsia
11	31	124.0	482	329	53	1.64	123	33.0	preclampsia
12	31	94.0	493	341	18	1.35	383	30.0	preclampsia
13	39	134.0	377	233	43	0.60	148	36.0	preclampsia

- 2) Se ingresa el datasets en la función metaMDS para hallar el escalamiento multi dimensional.

```
NMDS = metaMDS(dta)
```

```

Square root transformation
Wisconsin double standardization
Run 0 stress 0.1812327
Run 1 stress 0.2068479
Run 2 stress 0.1900273
Run 3 stress 0.1816361
... procrustes: rmse 0.007709973 max resid 0.07223566
Run 4 stress 0.2116889
Run 5 stress 0.212064
Run 6 stress 0.2003214
Run 7 stress 0.1920453
Run 8 stress 0.1975173
Run 9 stress 0.1824216
Run 10 stress 0.2064725
Run 11 stress 0.1872802
Run 12 stress 0.1811655
... New best solution
... procrustes: rmse 0.003175277 max resid 0.02803874
Run 13 stress 0.1977339
Run 14 stress 0.190094
Run 15 stress 0.191908
Run 16 stress 0.2092101
Run 17 stress 0.2055263
Run 18 stress 0.1901082
Run 19 stress 0.1812034
... procrustes: rmse 0.002725936 max resid 0.02452488
Run 20 stress 0.1824265

```

- 3) Se imprime la información de la variable NMDS

NMDS

```

Call:
metaMDS(comm = dta)

global Multidimensional Scaling using monoMDS

Data:   wisconsin(sqrt(dta))
Distance: bray

Dimensions: 2
Stress:   0.1811655
Stress type 1, weak ties
No convergent solutions - best solution after 20 tries
Scaling: centring, PC rotation, halfchange scaling
Species: expanded scores based on 'wisconsin(sqrt(dta))'

```

- 4) Se muestra la grafica de las distancias entre factores almacenadas en la variable NMDS como se observa en la Figura13.

`ordiplot(NMDS, type = "t")`

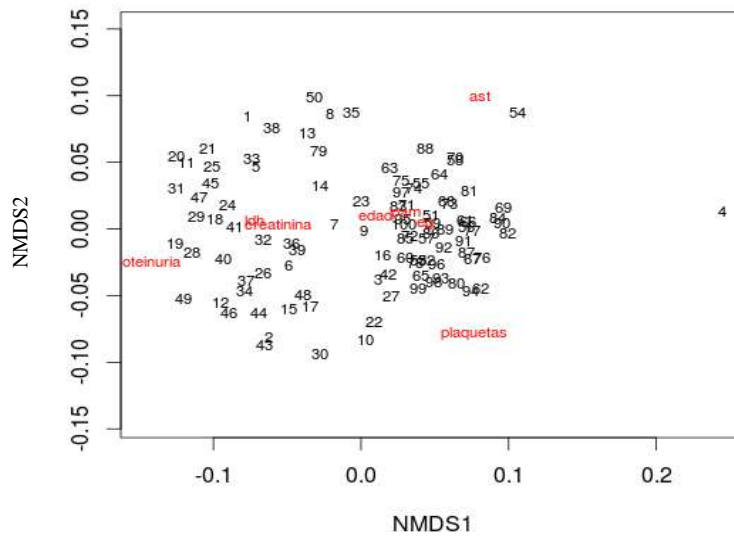


Figura 13. Distancia entre Factores

5) Se selecciona un conjunto objetos en 2 clases por medio de la bondad de ajuste y se grafica.

como se observa en la Figura14 y Figura15.

par (mfrow = c(1,2))

stressplot (NMDS)

plot (NMDS, display = 'sites', type = 't', main = 'Goodness of fit')

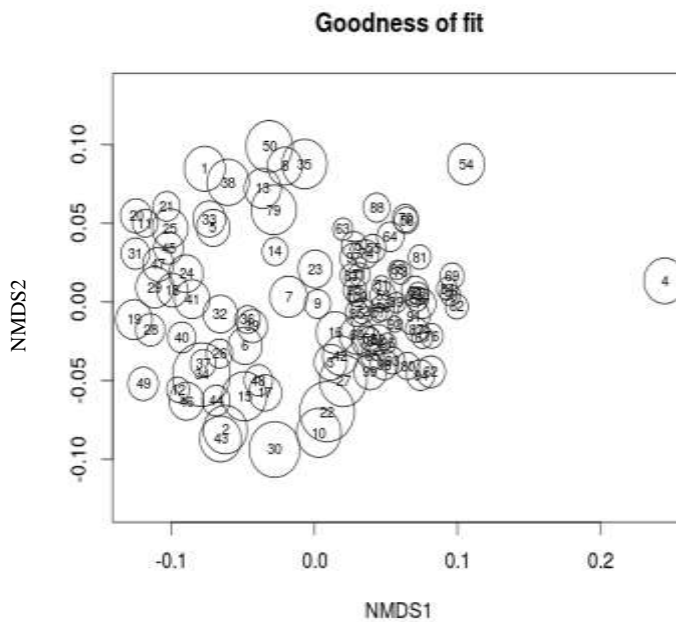


Figura 14. Distancia entre las especies de un dataset

De la anterior Figura14 se muestra la grafica de NMDS2 vs NMDS1 en donde es posible notar las proximidades existentes entre un conjunto de objetos similarmente parecidos en distancia y que a la vez muestra interdependencia con otro conjunto de datos que serían representados como sinpreeclampsia.

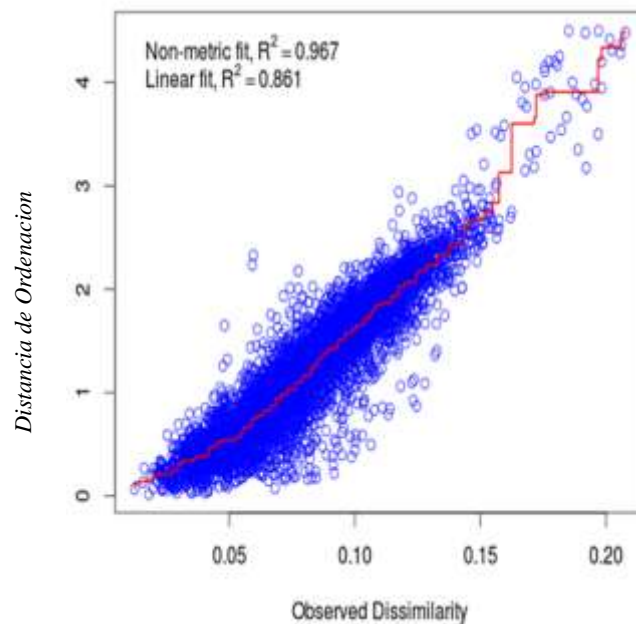


Figura 15. Medida de ajuste para puntos de escalamiento multidimensiona entre Distancia de Ordenacion vs Disimilitud de Factores

En la Figura 15 se muestra una la grafica de Distancia de Ordenación vs Disimilitud de Factores en donde la función stressplot es una envoltura para la función Shepard en el paquete MASS y traza las distancias de coordinación contra disimilitudes originales de los factores, y dibuja una línea de paso del ajuste no lineal. Además, se añade a la gráfica dos estadísticas de correlación que se extienden sobre la bondad del ajuste.

CAPÍTULO 6

6. PRUEBAS Y RESULTADOS

Una vez que cada registro de pacientes enfermos y sanos ha sido procesado, se extraen los valores y se construye un árbol de decisión como grupo de entrenamiento usando el concepto de entropía de información. Los datos de entrenamientos son un grupo de ejemplos que tiene un conjunto de atributos o características del ejemplo representando así la clase a la que pertenece cada muestra.

En cada nodo el árbol C5.0 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Este criterio es formalizado para la ganancia de información según la diferencia de entropía que resulta en la elección de un atributo para dividir los datos. El atributo con mayor ganancia de información normalizada se elige como parámetro de decisión.

En total se escogió un conjunto de 100 datos simulados bajo restricciones de variables, de los cuales 80 correspondiente al 75% de la muestra pertenecieron al conjunto de entrenamiento en donde se generaron las reglas de clasificación que constituyen la base para encontrar el factor predictor de la preeclampsia y los otros 20 corresponden al 25% de la muestra que sirvió como validación.

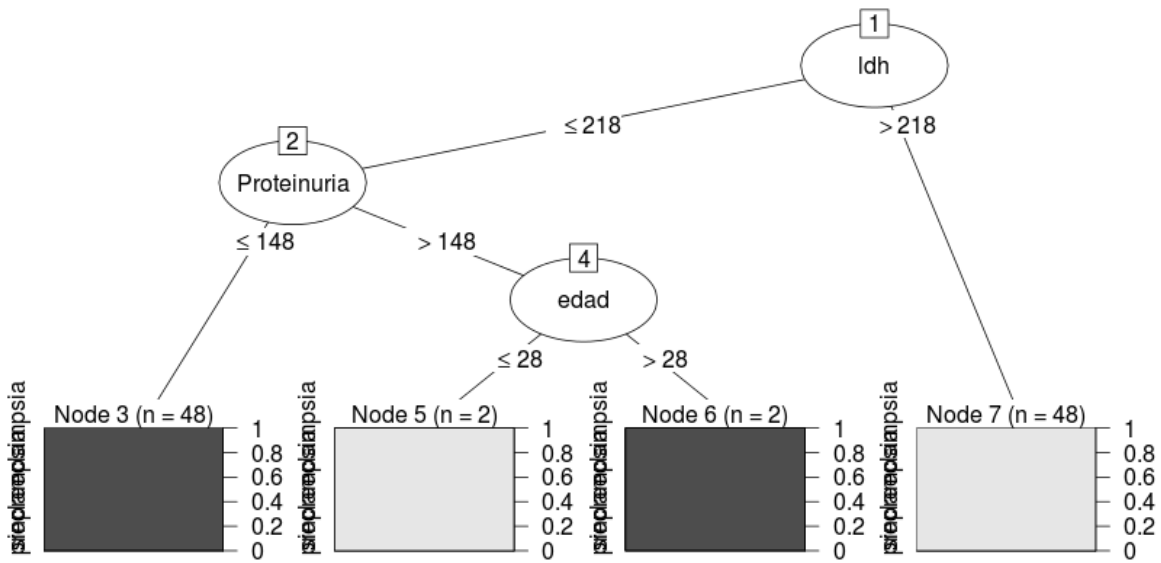


Figura 16. Arbol C5.0 - 3 nivel

Al granular el árbol C5.0 se encuentra que el segundo factor mas importante es la proteinuria seguido de la edad. Estableciendo los valores de la variable en la clasificación *Proteinuria* ≤ 150 y > 150 y los valores de *Edad* ≤ 28 y > 28 de (sano) sinpreeclampsia y (enfermo) preeclampsia mostrada en la Figura 16. Arbol C5.0 - 3 nivel .

```

Console
Call:
C5.0.default(x = datar[1:88, -18], y = datar[1:88, 18])

C5.0 [Release 2.07 GPL Edition]      Wed Feb 10 13:08:21 2016
-----

Class specified by attribute 'outcome'
Read 88 cases (18 attributes) from undefined.data
Decision tree:
ldh <= 218: sinpreeclampsia (44/2)
ldh > 218: preeclampsia (36)

Evaluation on training data (88 cases):

Decision Tree
-----
Size      Errors
-----
2         2( 2.5%) <<<

(a) (b) <-classified as
----
36  2  (a): class preeclampsia
42  42 (b): class sinpreeclampsia

Attribute usage:
100.00% ldh

```

Figura 17. Resultados del Entrenamiento C5.0

Para la cual la variable con más ganancia de información fue *ldh* (lactato deshidrogenasa) enzima catalizadora encontrada en muchos tejidos del cuerpo. Estableciendo los valores de la variable en la clasificación $ldh \leq 218$ y $ldh > 218$ de (sano) sinpreeclampsia y (enfermo) preeclampsia mostrada en la Figura 17. Resultados del Entrenamiento C5.0.

Según los resultados obtenidos con el algoritmo NMDS están subdivididos en 2 parte PCA Y PCoA los cuales realizan separación de los datos.

Utilizando la técnica PCA es posible observar dentro de un datasets que componentes son denominados principales como semuestra en la Figura 12. Separación Factores predictores. Donde se muestran que las variables que más impacto tienen en datasets son (*ldh*, proteinuria, plaquetas) y las cuales se alejan de las demás variables del modelo de datos.

Utilizando la técnica PCoA y recibiendo la matriz de distancias del PCA fue posible identificar qué factores pertenecían más a cada especie (preeclampsia, sin preeclampsia). Se observó que los factores como edad, pam (presión arterial media), eg (edad gestacional) no se separaban del conjunto de datos o especie sin preeclampsia como se muestra en la Figura 13. Distancia entre Factores.

Además realizando una Bondad de ajuste, es posible notar las proximidades existentes entre un conjunto de objetos similarmente parecidos en distancia y que a la vez muestra interdependencia con otro conjunto de datos son representados como sinpreeclampsia por lo cual se logró separar las dos especies o clases como se muestra en la Figura 14. Distancia entre las especies de un dataset.

Luego se utiliza la función stressplot de envoltura para la función Shepard en el paquete MASS. En esta se traza las distancias de coordinación contra disimilitudes originales, y dibuja una línea de paso del ajuste no lineal como se muestra en la siguiente Figura 15. Medida de ajuste para puntos de escalamiento multidimensional.

CAPÍTULO 7

7. CONCLUSIONES Y PERSPECTIVA

El sistema de clasificación de preeclampsia en mujeres embarazadas desarrollado, muestra cómo pueden implementar técnicas de inteligencia artificial en medicina enfocadas a las áreas ginecología, para el reconocimiento automático de enfermedades en las entidades de atención médicas de tercer nivel.

A partir de los resultados obtenidos en la implementación de este proyecto para el reconocimiento de la preeclampsia se puede concluir lo siguiente.

- La etapa de adquisición de la información juega un papel importante en el proceso de predicción de la preeclampsia. Puesto que depende en gran medida del extracto social al que pertenece la entidad médica a donde son atendidas estas mujeres lactancia.
- Es importante tener en cuenta, que se deben eliminar las variables que no aportan ganancia de información dentro del árbol entrenado con el algoritmo C5.0.
- La variable LDH resulto ser el factor más adecuado para la predicción de preeclampsia . Esto se puede reafirmar visualmente puesto que es la variable con mayor ganancia de información de las elegidas como predictoras (ver Figura 9. Arbol entrenado C5.0).
- La técnica PCA y NMDS fueron tenidas en cuenta para ver el comportamiento de las variables proximidades entre un conjunto de objetos similarmente parecidos en distancia, distancias de coordinación, estadísticas de correlación de las variables.

- Tanto en el algoritmo C5.0 como en las técnicas PCA y NMDS se visualizó que los factores LDH, Proteinuria y Edad son los factores más adecuados para la predicción de preeclampsia.
- La clasificación de preeclampsia aplicando el algoritmo de árboles de decisión C5.0 presentó un porcentaje de error del 2.5% dando un porcentaje de acierto del 97.5%. Una vez el árbol de decisión ha sido entrenado, el reconocimiento de la enfermedad es automático. Además según las pruebas de cribaje explicado por el autor (Morgado, 2008), utilizando las técnicas tradicionales (Mancuernas con presión arterial media y MAP con Factores de riesgo) para la predicción de preeclampsia, se encuentra una tasa de detección del 12% y 62.5% respectivamente.
- Comparando el porcentaje de acierto del modelo explicado en este documento, con el presentado por el autor (Neocleous, 2009), que aplicando la técnica de minería de datos, redes neuronales obtuvo un porcentaje de acierto del 83,6% de la preeclampsia y en la prueba de ajuste del 93,6%, por lo cual se observa una mayor eficiencia en el resultado del modelo propuesto en esta tesis.

ANEXOS

Barranquilla, 5 de diciembre de 2015

Doctor
Nicola Ambrosi
Director Medico
Clinica Portoazul
Barranquilla

Por medio de la presente, les solicito muy amablemente que me colaboren en el estudio que estoy realizando, titulado **"DISEÑO DE UNA HERRAMIENTA PARA ASISTENCIA MÉDICA Y PREDICCIÓN DE PREECLAMPSIA EN MUJERES EMBARAZADAS"** realizado por mi persona Roberto Porto, como requisito para la tesis de Maestría en Ingeniería de Sistemas y Computación, de la Universidad del Norte. Para esto se le solicitan datos no identificados de historias clínicas de pacientes embarazadas que presentaron preeclampsia y datos no identificados de mujeres embarazadas que no lo presentaron. Los datos a utilizar son los siguientes:

1. Edad de la mujer embarazada
2. Presión arterial sistólica
3. Presión arterial diastólica
4. Índice de lacto Deshidrogenasa
5. Índice de aparato Deshidrogenasa
6. Índice de Creatinina
7. Índice de plaquetas
8. Usted fuma o ha fumado durante el proceso de gestación?
9. Edad gestacional
10. En su familia se han presentado casos anteriores de Preeclampsia o Eclampsia
11. Cuál fue el método de concebir el embarazo?
12. Talla del paciente
13. Usted consume alcohol o ha consumido alcohol durante el proceso de gestación?
14. Peso de la paciente
15. Índice de masa Corporal

Cabe resaltar que solo se utilizarán para la investigación los datos mencionados anteriormente por tal motivo no es necesaria la información correspondiente a los datos personales.

Quedo atento a su respuesta. Muchas gracias por la atención prestada.


Roberto porto solano
roberhp@uinorte.edu.co
Tel: 3003590975




A QUIEN INTERESE

Dr. Nicola Ambrosi, Director Médico de la Clínica Portoazul y Especialista en Ginecología y Obstetricia, autoriza:

La presentación de la Tesis de Maestría titulada "DISEÑO DE UNA HERRAMIENTA PARA ASISTENCIA MÉDICA Y PREDICCIÓN DE PREECLAMPSIA EN MUJERES EMBARAZADAS" realizada por el ingeniero Roberto Porto, bajo mi dirección y supervisión, y que presenta para la obtención del grado de Maestría en ingeniería de sistemas y computación por la universidad del Norte.

Se expide la presente certificación a petición de la parte interesada a los veinte (20) días del mes de enero de dos mil diez y seis (2.016)



Dr. Nicola Ambrosi
Director Médico

Comité de Ética en investigación de la División
Ciencias de la Salud de la Universidad del Norte.

ACTA DE EVALUACION: N° 136

Fecha: 10 de Diciembre del 2015

Nombre Completo del Proyecto: "DISEÑO DE UNA HERRAMIENTA PARA ASISTENCIA MÉDICA Y PREDICCIÓN DE PREECLAMPSIA EN MUJERES EMBARAZADAS"

Nombre del Investigador y co-investigadores:

Roberto Porto (estudiante de maestría en Ingeniería de Sistemas y Computación)
Miguel Jimeno (profesor departamento de Ingeniería de Sistemas y Computación)
Eduardo Zurek (profesor departamento de Ingeniería de Sistemas y Computación)

Sitio en que se conduce o desarrolla la investigación: En la ciudad de Barranquilla.

Fecha en que fue sometido a consideración del comité: 10 de Diciembre del 2015

EL COMITÉ DE ÉTICA EN INVESTIGACIÓN EN EL ÁREA DE LA SALUD. Creado mediante Resolución rectoral N° 05 de Febrero 13 de 1995 en atención a la Resolución No. 008430 de 1993 del Ministerio de Salud como parte esencial para el funcionamiento de cualquier institución que realiza programas de investigación en humanos.

Conformado inicialmente por los siguientes miembros. Refrendado en el año 2005 con el objeto de ajustarse a estándares éticos y científicos de la investigación biomédica establecidos en la Declaración de Helsinki, Guías Operacionales para Comités de Ética de la OMS y las Guías para Buena Práctica Clínica del ICH.

Se acoge a las Buenas Prácticas Clínicas del ICH de acuerdo a la normativa vigente, Resolución N° 2378 del Ministerio de Protección Social, Declaración de Helsinki versión 2013 y guías operativas de OMS, Informe Belmont.

El comité de ética en investigación en el Área de la Salud Universidad del Norte certifica que:

1. Sus miembros revisaron los siguientes documentos del protocolo en referencia:

- Carta de presentación del proyecto generada por el Investigador
- Copia del proyecto completo de investigación
- Resumen ejecutivo
- Hojas de vida de los Investigadores

2. El presente proyecto fue evaluado por los siguientes miembros:

- Enf. GLORIA VISBAL ILLERA
Profesión: Enfermera, Mg. Bioética
Cargo en el Comité de Ética: Presidenta y Representante de Profesores
- Dr. RAFAEL TUESCA MOLINA
Profesión: MD. Phd. en Salud Pública
Cargo en el Comité de Ética: Representante Científico
- Dr. DIMAS BADEL MERLANO
Profesión: MD. Especialista en Bioética
Cargo en el Comité de Ética: Especialista en Bioética
- Dr. ROBERTO SOJO GONZÁLEZ
Profesión: Administrador de empresas
Cargo en el Comité de Ética: Representante de la Comunidad (Suplente)
- Q.F. MICHAEL MACIAS
Profesión: Químico Farmacéutico
Cargo en el Comité de Ética: Representante experto en Farmacia Química (Suplente)
- Dra. VIRIDIANA MOLINARES HASSAN
Profesión: Abogada
Cargo en el Comité de Ética: Representante No Científica (Suplente)
- Ing. PEDRO VILLALBA AMARIS
Profesión: Ingeniero Mecánico, Phd Ingeniero Biomédico
Cargo en el Comité de Ética: Representante Científico (Suplente)

3. El Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte establece que el número de miembros para que haya quórum es cinco (5), y se encuentra constituido por los siguientes miembros:

- Dr. HERNANDO BAQUERO LATORRE
Profesión: MD. Pediatra y Neonatólogo
Cargo en el Comité de Ética: Representante Científico
- Dra. OLGA HOYOS DE LOS RÍOS
Profesión: Phd en Psicología
Cargo en el Comité de Ética: Representante de Profesores
- Dra. SILVIA GLORIA DE VIVO
Profesión: Abogada
Cargo en el Comité de Ética: Representante No Científica
- Dr. RAFAEL TUESCA MOLINA
Profesión: MD. Phd. en Salud Pública
Cargo en el Comité de Ética: Representante Científico
- Dr. DIMAS BADEL MERLANO
Profesión: MD. Especialista en Bioética
Cargo en el Comité de Ética: Especialista en Bioética
- Enf. GLORIA VISBAL ILLERA
Profesión: Enfermera, Mg. Bioética
Cargo en el Comité de Ética: Presidenta y Representante de Profesores
- Dra. LOURDES MARTÍNEZ
Profesión: Administradora de empresas
Cargo en el Comité de Ética: Representante de la Comunidad

- Q.F. RICARDO AVILA
Profesión: Químico Farmacéutico
Cargo en el Comité de Ética: Representante experto en Farmacia Química
- Dra. NELLY LECOMPTÉ BELTRAN
Profesión: MD. Pediatra
Cargo en el Comité de Ética: Representante Científico (Suplente)
- Ing. JAIME GARCÍA OROZCO
Profesión: Ingeniero Mecánico
Cargo en el Comité de Ética: Representante de la Comunidad (Suplente)
- Dr. ROBERTO SOJO GONZÁLEZ
Profesión: Administrador de empresas
Cargo en el Comité de Ética: Representante de la Comunidad (Suplente)
- Dr. JORGE LUIS ACOSTA REYES
Profesión: MD. Mg. Ciencias Clínicas
Cargo en el Comité de Ética: Miembro - Representante Científico (Suplente)
- Dr. JEAN DAVID POLO VARGAS
Profesión: Psicólogo. Phd en comportamiento social y organizacional
Cargo en el Comité de Ética: Miembro - Representante de Profesores (Suplente)
- Enf. DIANA DÍAZ MASS
Profesión: Enfermera
Cargo en el Comité de Ética: Representante de Profesores (Suplente)
- Q.F. MICHAEL MACIAS
Profesión: Químico Farmacéutico
Cargo en el Comité de Ética: Representante experto en Farmacia Química (Suplente)
- Dra. VIRIDIANA MOLINARES HASSAN
Profesión: Abogada
Cargo en el Comité de Ética: Representante No Científica (Suplente)
- Ing. PEDRO VILLALBA AMARIS
Profesión: Ingeniero Mecánico. Phd Ingeniero Biomédico
Cargo en el Comité de Ética: Representante Científico (Suplente)

El Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte, se encuentra ubicado en la Universidad del Norte, KM 5 vía a Puerto Colombia. Primer piso Bloque F.

Contactos:

Correo electrónico: comite_eticauninorte@uninorte.edu.co

Página Web: www.uninorte.edu.co/divisiones/salud/comite_etica

Teléfono: 3509280 – 3509509 Ext. 3493

4. el comité considero que el presente estudio:

a. Es válido desde el punto de vista ético. La investigación se ajusta a los estándares de la buena práctica clínica.

5. El Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte informara inmediatamente a las directivas institucionales:

a. Eventos que son de notificación obligatoria por parte del investigador al comité de ética.

b. Cualquier cambio o modificación a este proyecto que haya sido revisado y aprobado por este comité.

● Km. 5 vía a Puerto Colombia • Apartado Aéreo 1559 - 51820 • Computador PSX: 3509500 • Fax: (05) 3508852 • Barranquilla, Colombia • www.uninorte.edu.co

6. El Comité informara inmediatamente a las directivas, toda información que reciba acerca de:

- a. Lesiones o daños a sujetos humanos con motivo de su participación en la investigación problemas imprevistos que involucren riesgos para los sujetos u otras personas cuando aplique.
- b. Cualquier cambio o modificación a este proyecto que haya sido revisado y aprobado por este comité.

7. Cuando el Protocolo es aprobado por el Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte, será por un periodo de un (1) año a partir de la fecha de su aprobación; según Guías Operativas CE_versión 18 ENERO 29 de 2015 literal *seguimiento a estudios aprobados el comité de ética en investigación.*

8. el Investigador principal deberá:


- a. Informar cualquier cambio que se proponga a introducir en el proyecto. Estos cambios no podrán ejecutarse sin la aprobación previa del COMITÉ DE ÉTICA EN INVESTIGACIÓN EN EL AREA DE SALUD DE LA UNIVERSIDAD DEL NORTE. Si estos son necesarios para minimizar o suprimir un peligro inminente o un riesgo grave para los sujetos que participan en la investigación deben ser notificados al comité de ética tan pronto sea posible cuando aplique.
- b. Notificar cualquier situación imprevista que implica algún riesgo para los sujetos o la comunidad o el medio en el cual se lleva a cabo el estudio cuando aplique.
- c. Informar la terminación prematura o suspensión del proyecto explicando causas y razones.
- d. Presentar a este comité un informe cuando haya transcurrido un año, contado a partir de la aprobación del proyecto. Los proyectos con duración mayor a un año, serán reevaluados a partir del primer informe entregado.
- e. Todos los proyectos deben entregar al finalizar un informe final de cierre del estudio, firmado por el investigador responsable.

9. Concepto del Comité de Ética

a. En reunión del Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte, realizada el 10 de Diciembre 2015, legalizada según acta No. 136, el consenso de sus miembros aprueba el proyecto de investigación en referencia.

Atentamente,


Nombre: GLORIA VISBAL ILLERA
Titulo: Enfermera, Mg. Bioética
Cargo: Presidenta Comité De Ética en Investigación del Área de la Salud
de la Universidad del Norte.

**UNIVERSIDAD DEL NORTE**
Comité de Ética en Investigación
en el Área de la Salud.

ENTREGADO. D. 5 MAR. 2016

BIBLIOGRAFÍA

- Agrawal, G. (2013). Optimization of C4.5 Decision Tree Algorithm for Data Mining Application. *International Journal of Emerging Technology and Advanced Engineering*, 5.
- Bagchi, A. (1997). RID3: An ID3-Like Algorithm for Real Data. *Information Sciences*, 290.
- Benzécri, J. (1980). donnees, L'analyse des. *Inra*, 31.
- Breiman, L. F. (1984). Classification and Regression Trees. *CRC press*.
- Cifuentes, R. (2010). *Preeclampsia-eclampsia diagnóstico y manejo*. Bogota: Altavoz.
- Fripp, G. (2014). Understanding Perceptual Maps for Marketing. *University of Sydney*, 11.
- Legendre, P. (1998). *Numerical Ecology* (1 ed.). Elsevier.
- morgado, R. (2008). *Metodos de predicción en la preeclampsia*. DOSSIER.
- Neocleous, C. K. (2009). *Neural networks to estimate the risk for preeclampsia occurrence*. Londres: IEEE.
- Quinlan, J. R. (1986). *Induction od decision trees*. Boston: Machine learning.
- Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.
- Quinlan, J. R. (1996). *Improved use of continuous attributes in c4.5*. Sydney: Journal of Artificial intelligence research .
- Ruggieri, S. (2002). Efficient C4.5. *IEEE*, 438- 444.
- Rulequest. (2012). *Is See5/C5.0 Better Than C4.5*. Obtenido de <http://rulequest.com/see5-comparison.html>.
- Bouchard, K. (2011). A new qualitative spatial recognition model based on Egenhofertopological approach using C4.5 algorithm: experiment and results.
- Bujlow, T. (2012). A method for classification of network traffic based on C5.0 Machine Learning Algorithm.
- Bujlow, T. (2012). Classification of HTTP traffic based on C5.0Machine Learning Algorithm.
- Cailliez, F. (1983). The analytical solution of the additive constant problem.
- Cambria, E. S. (2013). Semantic multi-dimensional scaling for open-domain sentiment analysis.

- Chi, Z. (1996). Handwritten Digit Recognition Using Combined Id3-Derived Fuzzy Rules And Markov Chains.
- Chiang, H.-J. (2011). A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm.
- Cleveland. (1979). Robust locally weighted regression and smoothing scatterplots.
- Cook EF, G. L. (1984). Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis.
- Dai, W. (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm.
- Deconinck, E. (2006). Classification tree models for the prediction of blood-brain barrier passage of drugs.
- Domínguez, M. J. (2012). ÁRBOLES DE CLASIFICACIÓN: UNA METODOLOGÍA PARA EL ANÁLISIS DE CRISIS BANCARIAS.
- Fahrmeir, L. a. (1984). Multivariate Statistische Verfahren, De Gruyter, Berlin.
- Foguet, J. M. (1988). Análisis multivariante: análisis de componentes principales.
- Fripp, G. (2014). Understanding Perceptual Maps for Marketing.
- González Martín, P. (2002). UNA APLICACIÓN DEL ANÁLISIS DE COMPONENTES PRINCIPALES EN EL ÁREA EDUCATIVA.
- Gower, J. (2005). Principal Coordinates Analysis Encyclopedia of Biostatistics.
- Holland, S. M. (2008). NON-METRIC MULTIDIMENSIONAL SCALING (MDS).
- Honarkhah, M. a. (2010). Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling.
- Ichihashi, H. (1995). Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning.
- Jin, C. (2012). A generalized fuzzy ID3 algorithm using generalized information entropy.
- Jinyu Guo, Y. L. (2010). Batch Process Monitoring Based on Multilinear Principal Component Analysis.
- Kale, A. (2015). Automated Menu Planning Algorithm for Children: Food Recommendation by Dietary Management System using ID3 for Indian Food Database.
- Lara Torralbo, J. A. (2008). MODELO PARA LA COMPARACIÓN DE DATOS POSTUROGRÁFICOS ESTRUCTURALMENTE COMPLEJOS.

- Legendre & Gower. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*.
- Lewis, R. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis*.
- Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices.
- Mak, B. (2000). Rule extraction from expert heuristics: A comparative study of rough sets with neural networks and ID3.
- Mantas, C. J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data.
- Mehta, S. (2015). Optimization of C5.0 Classifier using Bayesian Theory.
- Nie, B. (2009). Crowds' classification using hierarchical cluster, rough sets, principal component analysis and its combination.
- Niu, Z. (2009). Auto-Recognizing DBMS Workload Based on C5.0 Algorithm.
- Pandya, R. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning.
- Pang, S.-I. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks.
- Pashaei, E. (2015). Improving Medical Diagnosis Reliability Using Boosted C5.0 Decision Tree empowered by Particle Swarm Optimization.
- Polat, K. (2009). A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems.
- Rokach, L. (2008). *Data mining with decision trees: theory and applications*.
- Shao, X. (2000). Application of ID3 algorithm in knowledge acquisition for tolerance design.
- Shih Yin Ooi, A. B. (2016). Image-based handwritten signature verification using hybrid methods of discrete Radon transform, principal component analysis and probabilistic neural network.
- Smith, L. I. (2002). A tutorial on Principal Components Analysis.
- Timofeev, R. ... (2004). Classification and regression trees (cart). theory and applications. Master thesis, CASE - Center of Applied Statistics and Economics.
- Xiaogang, X. &. (2008). A Fault Detection Method Using Multi-Scale Kernel Principal Component Analysis.
- Xiaohu, W. (2012). An Application of Decision Tree Based on ID3.

Yang, R. (2012). DACIDR: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16S rRNA sequences.

PREECLAMPSIA ECLAMPSIA. (2007). *Via Cátedra de Medicina*.

Alejandro Bautista. (2012). Hipertensión arterial asociada con el embarazo. 1.

Alfaro, R. J. (s.f.). PREDICCIÓN DE PREECLAMPSIA - FACTORES DE RIESGOS. 7.

Banzhaf, J. (1965). Weighted Voting Doesn't: A mathematical Analysis. *Rutgers Law Review*, 317-343.

Cifuentes R. (2009). Preeclampsia-eclampsia. Diagnóstico y manejo En: Ginecología y Obstetricia Basadas en las Nuevas Evidencias . *Distribuna*, 345-348.

Cifuentes R. (2006). Hipertensión arterial y embarazo. En: Obstetricia de alto riesgo. *Distribuna*, 447-484.

Contreras F, B. M. (2003). Nuevos aspectos en el tratamiento de la pre-eclampsia y eclampsia. *Venez Farmacol Terap*, 1-23.

Dra. Verónica Natalia Joerin, D. L. (2007). PREECLAMPSIA ECLAMPSIA. *Via Cátedra de Medicina*.

Dustin T. Dunsmuir, B. A. (2014). *Development of mHealth Applications for Pre-Eclampsia Triage*.

Etchegaray Adolfo, S. M. (2012). Predicción de preeclampsia en el primer trimestre- validación prospectiva preliminar de un método de screening combinado.

Garovic, V. D. (2014). The Role of Angiogenic Factors in the Prediction and Diagnosis of Preeclampsia Superimposed on Chronic Hypertension. *Editorial Commentary*, 740-746.

Gillies, D. (1959). Solutions to general non-zero sum games. En A. Tucker, & R. Luce, *Contributions to the Theory of Games IV* (págs. 95-109). Princeton: Princeton University Press.

Harskamp RE., Z. G. (2007). Preeclampsia: At Risk for Remote Cardiovascular Disease. *Am J Med Sci*, 334:291-295.

Kelton, W. D., Smith, J. S., Sturrock, D. T., & Verbraeck, A. (2011). *Simio and Simulation: Modeling*. Simio LLC.

L.M.Fu. (1994). Neural networks in computer intelligence. New York.

Owen, G. (1995). A Value for Non-Transferable Utility Games. *Journal of Game Theory*, 95-109.

Raymundo. (2015). Prueba T de Student para datos relacionados. Obtenido de http://www.ray-design.com.mx/psicoparaest/index.php?option=com_content&view=article&id=232:t-student-dr&catid=52:pruebaspara&Itemid=61

- Redondo, C. N.-C. (2011). La importancia de la ecografía a las 11+0 a 13+6 semanas de embarazo.
- Riera, V. A. (24 de 3 de 2005). *spss*. Obtenido de http://www.ub.edu/aplica_infor/spss/cap4-7.htm
- Rivera, H. M. (2002). *Probabilidad y Estadística*. Obtenido de niversidad Nacional de Colombia: <http://www.virtual.unal.edu.co/cursos/ciencias/2001065/>
- S.Y. Leemaqz, G. D. (2013). Tiered Prediction System for Preeclampsia: an integrative application of multiple models .
- Serrano NC, S. (2005). nfluencia de los factores genéticos y medioambientales en la susceptibilidad para desarrollar preeclampsia. *MEDUNAB*.
- Sibai B, D. G. (2005). Pre-eclampsia. *Lancet*, 785-99.
- Vázquez, C. (21 de octubre de 2011). *EROSKI CONSUMER*. Obtenido de <http://www.consumer.es/web/es/bebe/embarazo/sintomas/2011/10/20/204305.php>
- Verlohren, S. (2012). Angiogenic growth factors in the diagnosis and prediction of pre-eclampsia.