Ediciones pensando el futuro
Editions thinking the future
Informática 2013
XV CONVENCIÓN Y FERIA INTERNATIONAL
15TH INTERNATIONAL CONVENTION AND FAIR
www.informaticahabana.com
www.informaticahabana.cu
ISBN: 978-959-7213-02-4

SLD133 CLASIFICADORES SUPERVISADOS PARA EL ANÁLISIS PREDICTIVO DE MUERTE Y SOBREVIDA MATERNA

SLD133 SUPERVISED CLASSIFIERS FOR PREDICTIVE ANALYSIS OF MATERNAL DEATH AND SURVIVAL

Pilar Vanessa Hidalgo León[1]

1 UPV-UNSAAC, Perú, pilarv_hidalgoleon@hotmail.com, San Jerónimo, Cusco-Perú

RESUMEN: *El presente trabajo se basa en el análisis de los clasificadores supervisados que puedan generar resultados aceptables para la predicción de la muerte y sobrevida materna, según características de pacientes complicadas durante su gestación, determinadas por los expertos salubristas. Muestra la problemática del tema, la justificación de la investigación y porque el alto índice de mortalidad materna en el país tiene una gran significancia en el nivel de desarrollo y la vulneración de los derechos humanos. Se describe la metodología del desarrollo, la descripción de la muestra y sus características además los instrumentos utilizados para el procesamiento de los datos. Luego los resultados de la investigación, y la conclusión a la que se llega después de la evaluación de cada clasificador y entre ellos el que mejores resultados arroja. Los histogramas acerca de cada atributo de las pacientes, y su inclusión en la muestra. Muestra también el parámetro determinante para su correcta clasificación. Además de los resultados comparativos entre cada tipo de clasificador dentro de la familia a la que pertenece. Al final se propone después de identificado el clasificador, implementar en R Project, el algoritmo de Naive-Bayes con estimador de Núcleo activado (KERNEL=TRUE), para su implementación en el sistema sanitario contribuyendo a la toma de decisiones certera y respaldada para los profesionales de la salud. En conclusión se encontró un clasificador supervisado que responde positivamente a dar cambio y mejora de la problemática que abarca a la sobrevida materna a pesar de sus complicaciones.*

Palabras Clave: Mortalidad materna, clasificadores supervisados, Redes bayesianas, Aprendizaje supervisado, Redes Neuronales, Algoritmos Basados en Distancias

ABSTRACT: *This paperis based onanalysis ofsupervised classifiersthat can generateacceptable resultsfor the predictionof death andmaternalsurvival, according to characteristics ofcomplicatedpatientsduringpregnancy, expertsidentified byhealth professionals. Displaysthe problematicissueofthe justificationof the investigationand thatthe high rate ofmaternal mortalityin the countryhas a greatsignificanceinthe level ofdevelopment andhuman rightsviolations. We describe thedevelopment methodology, the descriptionof the samplecharacteristicsandalso theinstruments usedfordataprocessing. Thenthe resultsof the investigationandtheconclusionare reached afterevaluating eachclassifierand amongthebest performingthrows. The histogramson each attributeof thepatients, and their inclusion inthe sample. It also shows theparameter forcorrect classification. Besidesyoucomparativeresultsbetween eachtype ofclassifierwithin the familyto which it belongs. In the endit is proposedafterthe classifieridentified, implemented inrprojectalgorithmnaive-bayes withkernelestimatoron (kernel =true), for implementationin the health systemcontributing todecision makingaccurateand supportedforhealth professionals. In conclusionwe found asupervisedclassifierwhich responds positively tomaking changeandimproving theproblematiccoveringmaternal survivaldespiteits complications.*

KeyWords: Maternal mortality, supervised classifiers, Bayesian networks, supervised learning,Neural Networks, Instance-based Algorithms.

1. INTRODUCCIÓN

El objetivo en el uso de clasificadores supervisados [1], es construir modelos que optimicen un criterio de rendimiento, utilizando datos o experiencia previa. En ausencia de la experiencia humana, para resolver una disyuntiva que requiere explicación precisa, los sistemas implementados por modelos clasificadores han sido parte importante en la toma de decisiones. A parte, cuando este problema requiere prontitud por su naturaleza, los clasificadores transforman los datos en conocimiento y aportan aplicaciones exitosas.

En el caso de factores de riesgo para la salud materna, existen estudios estadísticos y aplicaciones salubristas para determinarla, mas no integrados simultáneamente como parte de una probabilidad clasificatoria como modelo.

Por ello, este estudio determinará, el clasificador supervisado más eficiente en tiempo y resultado que establezca la brecha de clases entre pacientes gestantes complicadas durante su embarazo que pueden llegar a presentar síntomas fatales y las que no, así apoyar al personal de salud a tomar la decisión más optima y a prevenir futuras alzas en el índice de mortalidad de su comunidad.

Los problemas que generan el alza de este indicador ya son conocidos y puestos en valor en esta investigación:

"Hoy en día existe suficiente evidencia que demuestra que las principales causas de la muerte materna son la hemorragia posparto, la preclampsia o la sepsis y los problemas relacionados con la presentación del feto. Asimismo, sabemos cuáles son las medidas más eficaces y seguras para tratar estas emergencias obstétricas. Para poder aplicarlas, es necesario que la gestante acceda a un establecimiento de salud con capacidad resolutiva, pero lamentablemente muchas mujeres indígenas no acuden a este servicio por diversas razones, tanto relacionadas con las características geográficas, económicas, sociales y culturales de sus grupos poblacionales, como por las deficiencias del propio sistema de salud.

En los últimos años se han hecho muchos esfuerzos para revertir esta situación, tanto mediante proyectos promovidos por el estado como ejecutados por organismos no gubernamentales de desarrollo.

Estos esfuerzos han tenido, sin embargo, resultados desiguales debido principalmente a la poca adecuación de los proyectos al contexto geográfico y de infraestructura en el que vive gran parte de la población indígena, a sus dificultades económicas para acceder al servicio, su cultura, sus propios conceptos de salud y enfermedad, y su sistema de salud."[2].

2. CONTENIDO

2.1. Planteamiento del problema

¿Cuáles son los clasificadores supervisados que predicen la muerte o la sobrevida materna con mayor efectividad?

2.1.1. Específicos:

- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Redes neuronales en relación a los datos de mortalidad materna?
- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Redes Bayesianas en relación a los datos de mortalidad materna?
- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Regresión Logística en relación a los datos de mortalidad materna?
- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Arboles de Decisión en relación a los datos muerte y sobrevida materna?
- ¿Cuál es la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Algoritmos Basados en Distancias en relación a los datos muerte y sobrevida materna?

Se analizó los clasificadores supervisados como, arboles de clasificación, redes bayesianas, redes neuronales (perceptrón) y regresión logística.

Determinando mediante la herramienta Weka, la sensibilidad y certeza más cercana de cada uno de estos algoritmos, y cuya conclusión sugerirá el más eficiente.

En la actualidad la problemática en mortalidad materna es un indicador determinante de desarrollo en los países Latinoamericanos. Siendo no solo un indicador de pobreza y desigualdad sino de vulnerabilidad de los derechos de la mujer. [3]

2.2. Limitaciones de la Investigación

Los datos recolectados para este estudio con respecto a pacientes fallecidas que tuvieron complicaciones durante el embarazo fueron 48, pues fueron las que se registraron de manera legible y legal en los archivos de la Dirección Regional de Salud Cusco, esto hace que la muestra no pueda ser nutrida con mayor diversidad de datos.

Una limitación también es la poca investigación acerca del tema relacionado con el uso de clasificadores supervisados, pues existen muchas otras correspondientes a diagnósticos médicos que tienen similitud con la muestra pero ninguna que se relacione directamente.

La relevancia de los datos se limito a los antecedentes sobre estudios en mortalidad materna (edad, estado civil, analfabeta, ocupación, procedencia, anticoncepción, entorno (estrato social), controles pre-natales, ubicación domiciliaria, tiempo de demora en atención, atención profesional, antecedentes familiares, espacio intergenésico (en años), paridad (#de hijos), complicaciones no tratadas, fallecimiento), por lo cual se conservó el anonimato de cada paciente.

Se inició con un proceso de recolección de datos y conservación de la legitimidad de las historias clínicas proporcionadas por los hospitales de la región.[3]

2.3. Objetivos

- Determinar el clasificador supervisado que brinde mejores resultados para el análisis predictivo de muerte y sobrevida materna.

2.3.1. Objetivos específicos:

- Determinar la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Redes neuronales en relación a los datos de muerte y sobrevida materna.
- Determinar la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Redes Bayesianas en relación a los datos de muerte y sobrevida materna.
- Determinar la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Regresión Logística en relación a los datos muerte y sobrevida materna.
- Determinar la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Arboles de Decisión en relación a los datos muerte y sobrevida materna.
- Determinar la especificidad, la clasificación correcta y el error absoluto y sensibilidad del clasificador supervisado Algoritmos basado en distancia y en relación a los datos muerte y sobrevida materna.

2.4. Hipótesis General.

Hi: Existen clasificadores supervisados que predicen la muerte o la sobrevida materna con efectividad

Ho: No existen clasificadores supervisados que predicen la muerte o la sobrevida materna con efectividad

Ha: Algunos clasificadores supervisados predicen la muerte o la sobrevida materna con efectividad

2.4.1. Hipótesis Específicas.

- Las Redes Bayesianas brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad, clasificación correcta, error absoluto y sensibilidad recomendada.
- Las Redes Neuronales brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad, clasificación correcta, error absoluto y sensibilidad recomendada.
- Los Arboles de Decisión brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad, clasificación correcta, error absoluto y sensibilidad recomendada.
- Los Algoritmos basados en Distancias brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad, clasificación correcta, error absoluto y sensibilidad recomendada.
- La Regresión Logística brinda al estudio de los datos en mortalidad y sobrevida materna una especificidad, clasificación correcta, error absoluto y sensibilidad recomendada.

2.5. Definición de Variables:

2.5.1. Variable principal:

Clasificadores supervisados

2.5.2. Variables Implicadas:

Tabla I: Variables implicadas

Variable	Clasificadores supervisados
Dimensión	Las Redes Neuronales, Algoritmos supervisados, Las Redes Bayesianas, Arboles de decisión, Regresión Logística, Algoritmos basados en instancias
Indicador/ Criterios de Medición	Clasificación correcta, Clasificación incorrecta, Sensibilidad,Especificidad, Tiempo de ejecución, Mean absoluteerror,Tiempo de ejecución, Kappa statistic, Root mean squared error, Relative absolute error, Root relative squared error
Clase	Numérica discreta
Instrumento	Weka 3.5.7, Explored

2.6. Metodología de investigación

Cuasi Experimental, Aplicada, Inductivo

2.6.1. Descripción de la muestra y método de recolección

Los datos recolectados en todas 100 historias clínicas de casos de sobrevida y de casos de muerte materna.

Las características de las gestantes son muy similares entre si y corresponden a la población de la ciudad del Cusco, del archivo en la Red Sur de la Dirección Regional de Salud sobre el control sanitario de mortalidad materna de la Región Cusco.

El número es limitado, pues las historias desde 1992 al 2011, no han sido redactadas ni conservadas en el mejor estado haciendo difícil la tarea de interpretar los datos suficientes para ser analizados.

Estas pacientes no fueron necesariamente atendidas desde el inicio en estos establecimientos, sino que debido a sus complicaciones durante el parto y e embarazo fueron derivadas a las capitales y luego a los establecimientos de mayor capacidad resolutiva para su atención.

Los datos de las historias clínicas que incluyeran:

- Edad
- Estado civil
- Analfabeta
- Ocupación
- Procedencia
- Anticoncepción
- Entorno (estrato social)
- Controles pre-natales
- Ubicación domiciliaria
- Tiempo de demora en atención
- Atención profesional
- Antecedentes familiares
- Espacio intergenésico (en años)
- Paridad (#de hijos)
- Complicaciones no tratadas
- Fallecimiento

Fueron las HC que se utilizaron para la muestra test. Entre 52 sobrevivientes y 48 fallecidas, ambos grupos con similares características, siendo factores determinantes: [4-8]

Ubicación domiciliaria /tiempo demora en atención
Controles > 2: n (49.0/2.0)
PRODECENCIA = rural: s (6.0/1.0)

2.7. Técnica e instrumentos de investigación

Se utilizó la herramienta Weka Explorer para la interpretación de los datos.
Las opciones de clasificación supervisada y los algoritmos que propone esta herramienta. [9-11]

Se evaluaron los siguientes clasificadores supervisados:

- Las Redes Neuronales
  - MultilayerPerceptron
  - RBFNetwork
- Las Redes Bayesianas
  - BayesNet
  - Bayes simple estimator
  - BMA bayes
  - Naive-Bayes
  - BayesNet Kernel
  - Naive-Bayes Discretizacion Supervisada
- Arboles de decisión
  - J48
  - DecisionTable
- Regresión Logística
  - MultiClassClassifier

  - Logistic
- Algoritmos basados en Distancias
  - IBK
  - LWL
  - KStar

Estos resultados fueron comparados con las reglas de clasificación que nos proporcionará los algoritmos de predicción inmediata:

- OneR
- ZeroR

2.8. Procedimiento de recolección de datos

Las Historias Clínicas se insertaron en un fichero CSV (delimitado por comas) cuya cabecera se trata de las etiquetas de cada atributo, y la última columna se refiere a la clase a la pertenecen.
Con respecto a los indicadores particulares en cada uno de los atributos de los sujetos de la clase tenemos los siguientes valores: Tabla II

- La calidad de la estructuración
- Comparar todas mediciones en cada clasificador por algoritmo.
- Evaluar los resultados.
- Interpretar los resultados.

2.10. Plan de análisis de la información

- Determinación de objetivos
- Preparación de datos
- Selección: Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.
- Pre procesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
- Transformación de datos: conversión de datos en un modelo analítico.
- Análisis de datos interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.
- Asimilación de conocimiento: aplicación del conocimiento descubierto
- Minería de datos: tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos. [12-14].

Tabla II: Valores de los atributos en las pacientes de la muestra

	NEGATIVO	POSITIVO	VALORES
Edad	Menor a 19 y mayor 35	Entre 19-35	14-48
Estado civil	Soltera	Pareja	Soltera-pareja
Analfabeta	Analfabeta	Primaria	Analfabeta-primaria-secundaria-superior
Ocupacion	No remunerada	Remunerada	Remunerada-no remunerada
Prodecencia	Rural	Urbana	Rural-urbana
Anticoncepcion	No	Si	Si-no
Entorno (estrato social)	Bajo	Medio-alto	Baja-alta
Controles	De 1 a 5	6 a mas	0-12
Ubicación domiciliaria/tiempo demora en atención	Más de 2 horas	Menos 3 del ee.ss	Menos de 1 hora, 1-2,3-5,6 a mas
Personal de atención profesional	No	Si	No-si
Antecedentes familiares	No	Si	No-si
Espacio intergenésico (en años)	Menos de 2 mayor a 4	Entre 2-4	Primera gesta,1-3,4-6, menos 1
Paridad (#de hijos)	Primípara o más de 4	Entre 2-4	0-10
Complicaciones no tratadas	Complicaciones antes y durante	Sin complicaciones	No-si
Fallecida	Si	No	No-si

2.11. Histogramas:

Cada Atributo es evaluado visualmente por los histogramas que arroja Weka (en total 15), por ejemplo con respecto a la EDAD de las pacientes de la muestra.

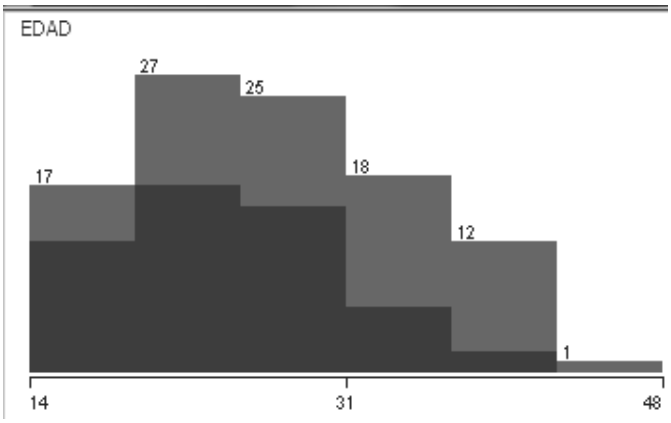


Gráfico 1: Histograma de la Edad de las pacientes de la muestra

Este histograma nos muestra el intervalo de edad de las pacientes de la muestra, los colores azul y rojo determinan la clase a la que pertenece cada intervalo. Podemos observar lo siguiente:

- Las pacientes del intervalo 14-20 años pertenecen en su mayoría a la clase "sobreviviente"
- Las pacientes del intervalo 21-31 años tienen un mayor porcentaje de sobrevida, coincide con el promedio de edad adecuado y de la muestra.
- El intervalo de paciente entre 32-40 años tiene mayor porcentaje de muerte.
- Las pacientes mayores a 40 años pertenecen a la clase "fallecida" en gran porcentaje.

2.12. Resultados por algoritmo testeado:

Los algoritmos usados para evaluar la base de datos en mortalidad materna dieron como resultado cifras continuas indicando los siguientes sucesos: Tabla 2.

- **Especificidad:** es la probabilidad de que pacientes complicadas y de riesgo pertenezcan a la clase Sobreviviente. Es decir los verdaderos negativos.

$$Especificidad = \frac{VN}{VN + FP}$$

- *Fracción de verdaderos negativos (FVN)*. Demuestra la cantidad de pacientes que realmente pertenecen a la clase Sobreviviente. Quiere decir que si el algoritmo estudiado tiene alto porcentaje de especificidad determina con gran éxito la probabilidad de sobrevida en pacientes complicadas durante su embarazo según los datos proporcionados en la ficha de antecedentes.
- **Clasificación correcta**: de la totalidad de datos, entre los que 52 que pertenecen a la clase Sobreviviente, y los 48 que pertenece a la clase Fallecida, determina dentro de cada clase cuantas instancias luego de la construcción del clasificador cuantas si pertenecen a la clase determinada.
- En el caso de pertenecer a la clase sobreviviente o a la fallecida de las 100 instancias cuantas fueron clasificadas correctamente.
- **Clasificación incorrecta:**del mismo modo la cantidad de instancias que no fueron clasificadas de manera correcta, son las que de manera supervisada se sabe que pertenecen a una u otra clase y fueron incluidas dentro de la cual no eran. Si el indicador emite un número mayor al 50% de la cantidad total de instancias, no se debe considerar como eficiente.
- **Sensibilidad:** es la capacidad del algoritmo de clasificar a las pacientes complicadas dentro de la clase Fallecidas. Es decir que si el clasificador tiene un alto porcentaje tiene mejor curva de corte y discernimiento entre los sujetos que pertenecen o no a la clase fallecida, es así que si la cifra de sensibilidad es del 90%, existe entonces esa probabilidad de que la paciente fallezca.

$$Sensibilidad = \frac{VP}{VP + FN}$$

- **Mean absolute error:** Se define **error absoluto** de una medida la diferencia entre el valor medio obtenido y el hallado en esa medida todo en valor absoluto.
- Entonces el promedio de error absoluto, es la suma de los errores absolutos de clasificación en cada uno de los sujetos llevados a promedio. El clasificador que arroje mayor cifra (mayor a 0.1) define un error de clasificación alto, por lo cual no se

debe considerar por sobre los que arrojen una cifra menor.

- **Tiempo de ejecución:** medido en segundos es la cantidad de tiempo que demora en construir la arquitectura del clasificador y en arrojar resultados.
- Puede que un clasificador se defina como eficiente si el tiempo que emplea en emitir resultados es menor a 5 segundos, aun así depende de los demás indicadores para valerse de esta característica.

- **Kappa statistic:** el Kappa statistic es la concordancia de comparación que tienen los observadores de clasificación. Quiere decir en una matriz de clasificación, el índice esperado entre el diagonal principal esperada (Xii elemento clasificado en la misma clase por ambos observadores) y el índice real luego de la clasificación efectuada por la arquitectura seleccionada (sea regresión lineal, backpropagation, Naive-bayes, etc.), es la diferencia en porcentaje de su lejanía a este valor.

- Si por ejemplo, la matriz esperada clasifica el valor en 25.00 y el resultado de la arquitectura es 26.7, la diferencia seria, 1.7 equivale al 90.32%. Entonces cuanto mas grande sea el porcentaje, estará más cerca de ser considerado eficiente.

- **Root mean squared error:** error cuadrático medio, es una medida de uso frecuente de las diferencias entre los valores pronosticados por un modelo o un estimador y los valores realmente observados. RMSD es una buena medida de precisión, pero sólo para comparar diferentes errores de predicción dentro de un conjunto de datos y no entre los diferentes, ya que es dependiente de la una escala muestra. Estas diferencias individuales también se denominan residuos, y la RMSD sirve para agregarlos en una sola medida de la capacidad de predicción.

- **Relative absolute error:** es el error relativo a cada característica de la clase, por ejemplo el error relativo de tener de Espacio Intergenésico 0.5 años y pertenecer o no a la clase fallecida, la clasificación indicaría que si pertenece, por ser el valor indicado para aquellas pacientes que están en peligro.
  En este caso el valor positivo para pertenecer al clase sobreviviente es de entre 2-4 años o primera gesta: 0.5 años incluido en la clase fallecido si el error entre los valores determinados por la clase y el valor ingresado es menor a 1.

- **Root relative squared error:** La **raíz relativa** *E* **de error al cuadrado** $_i$ de un programa individual *i* es evaluado por la ecuación:

$$Ei = \sqrt{\sum_{j-1}^{n} \frac{\left(P_{(ij)} - T_j\right)^2}{\sum_{j-1}^{n}\left(T_j - \overline{T}\right)^2}}$$

donde $P_{(ij)}$ es el valor predicho por el programa para el individuo *i j* muestra de casos (de los casos de la muestra *n),* T_i es el valor objetivo para *la* muestra *j* caso, y $\overline{T}$ está dada por la fórmula:

$$\overline{T} = \frac{1}{n}\sum_{j-1}^{n} T_j$$

Para un ajuste perfecto, el numerador es igual a 0 y $E_i = 0$. Así, el E_i índice varía de 0 a infinito, con el ideal que corresponde a 0.

Tabla III: Algoritmos de clasificación supervisada con mejores resultados por familia

INDICADORES	Reglas OneR	Redes Bayesianas NAÏVE BAYES KERNEL	Redes Neuronales RBFNetwork
Especificidad	0.836	0.91	0.9
Clasificación correcta	84	91	90
Clasificación incorrecta	16	9	10
Sensibilidad	0.868	0.914	0.9
Mean absolute error	0.16	0.1142	0.1468
Tiempo de ejecución	0.2	0.01	0.26
Kappa statistic	0.6759	0.819	0.7997
Root mean squared error	0.4	0.2737	0.3032
Relative absolute error	32.03%	22.86%	29.39%
Root relative squared error	80.01%	54.74%	60.64%

Tabla IV: Algoritmos de clasificación supervisada con mejores resultados por familia

INDICADORES	Arboles de Decisión J48	Algoritmos de Distancia LWL	Regresión Logística Logistic
Especificidad	0.9	0.889	0.82
Clasificación correcta	90	89	82
Clasificación incorrecta	10	11	18
Sensibilidad	0.9	0.902	0.82
Mean absolute error	0.1174	0.168	0.1753
Tiempo de ejecución	0.02	0	0.81
Kappa statistic	0.7994	0.6388	0.6388
Root mean squared error	0.299	0.3019	0.4157
Relative absolute error	23.75%	33.64%	35.10%
Root relative squared error	59.80%	60.39%	83.15%

2.13. Descripción de la metodología propuesta

Se propone entonces que evaluando cada uno de los indicadores más importantes en la construcción del clasificador, se descarten aquellos que no cumplen las siguientes características:

- Especificidad > 90%
- Clasificación correcta > 90 instancias
- Clasificación incorrecta < 10 instancias
- Sensibilidad >90%
- Mean absolute error < 0.1 ideal
- Kappa statistic >0.79, >0.9 ideal
- Root mean squared error < 0.3, <1 ideal
- Relative absolute error <25%, <1 ideal
- Root relative squared <50%, 0 ideal

El clasificador que cumpla con estas especificaciones, se considera como óptimo para la integración en un sistema que evaluará la base de datos que contengan los datos de las pacientes a comparar con un registro nuevo de entrada.

2.13.1. Etapa de Evaluación:

Cada uno de los clasificadores estudiados, que analizaron la muestra de pacientes, arrojaron los indicadores que muestra la tabla, en nivel de rendimiento y buenos resultados, aceptables para el estudio, la arquitectura que lleva los porcentajes más óptimos dentro de los parámetros, es el algoritmo de Naive bayes con estimador de núcleo.

Descartando así el resto de arquitecturas como apropiadas para este tipo de muestras. Además de visualizar de manera más clara y correcta los resultados comunes entre cada una de las arquitecturas dejando atrás aquel paradigma que incluye a la regresión logística como la más adecuada a la hora de realizar diagnósticos preventivos en salud.

Kernel estimator, al ser un método de construcción no paramétrico, es más flexible que los clasificadores que incluyen parámetros. Dividiendo la muestra en 40 grupos en la variable edad, paridad 15, controles 17, para la exploración descubre nuevas alternativas de clasificación en los atributos de la clase, mostrándolos como relevantes.

El problema más resaltante es que requiere mayor tamaño de muestra para mantener estos resultados, ya que utilizado el agrupamiento para la clasificación (clustering) que podría ser una desventaja si de clasificadores supervisados hablamos. Esto se denomina the curse of dimensionality, pues dependen de la elección de un parámetro suavizado, en este caso los valores la edad, número d controles, y número de hijos, transformándola en no objetiva.

Cuando Naive-Bayes actúa con la estimación no paramétrica, las estructuras que se construyen a partir de esta arquitectura se obtienen a partir de un árbol formado con variables potencialmente predictoras multidimensionales.[15-17]

3. CONCLUSIONES

Con respecto a la respuesta de las hipótesis podemos afirmar lo siguiente:

- Encontramos que las arquitecturas propuestas por esta memoria, totas conservan una efectividad bastante aceptable e cada uno de sus indicadores que en conjunto hacen un 80% de eficacia. Siendo el más óptimo el de Naive-Bayer Kernel.

- Descarta estas hipótesis luego del trabajo realizado.

- Las Redes Bayesianas brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 91%,

clasificación correcta 91%., error absoluto 0.1142, y sensibilidad 0.914 recomendada.

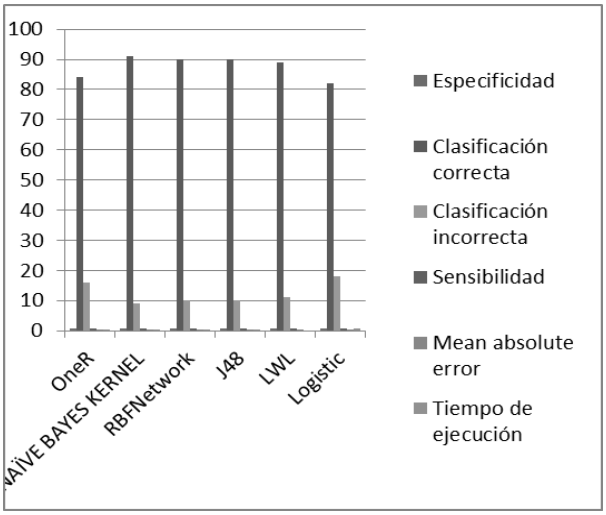
- Las Redes Neuronales brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 90%, clasificación correcta 90%, error absoluto 0.1468 y sensibilidad 90% recomendada.

- Los Arboles de Decisión brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 90%, clasificación correcta 90%, error absoluto 0.1174 y sensibilidad 90% recomendada.

- Los Algoritmos basados en Distancias brindan al estudio de los datos en mortalidad y sobrevida materna una especificidad 88.9%, clasificación correcta 89%, error absoluto 0.168 y sensibilidad 90.2% recomendada.

- La Regresión Logística brinda al estudio de los datos en mortalidad y sobrevida materna una especificidad 82%, clasificación correcta 82%, error absoluto 0.1753 y sensibilidad 82% recomendada.

Gráfico 2: Indicadores estadísticos de todos los algoritmos de clasificación supervisada



3.1. RECOMENDACIONES

- Se recomienda usar datos discretizados.
- Agrupar los datos de la manera propuesta en la Tabla 1, así podremos respetar los parámetros acerca de mortalidad materna que establecen los expertos salubristas.
- Es importante, comparar los resultados con la regla de decisión OneR para próximos experimentos, pues nos da la mejor noción de veracidad y efectividad a la hora de analizar la información.
- Se pretende implementar un sistema en R Project, para ingresar nuevos registros y que lleve en la memoria la base de datos recolectada a través de este trabajo.
- Para ayudar a la muestra numeraria que requiere Naive- Bayes Kernel, es necesario ingresar los registros de las muertes maternas y complicaciones diarias de las gestantes a nivel Nacional en la base de datos.

4. AGRADECIMIENTOS

- Agradezco al Dr. Basilio Araujo y al Dr. YosuYuramendi, por su apoyo y perseverancia para la culminación de este proyecto de fin de master.
- Al personal de salud que proporciono la información clínica que se usó en la investigación.
- A mi familia por el tiempo que no les pude dedicar y por su enorme comprensión y amor incondicional. A Alex por su amor, paciencia y comprensión

5. REFERENCIAS BIBLIOGRAFICAS

1. Clasificadores supervisados: el objetivo es obtener un modelo clasificatorio valido para permitir tratar casos futuros. Araujo, B. S. *Aprendizaje Automatico: conceptos básicos y avanzados.* Madrid, España: Pearson Prentice Hall, 2006.

2. Cordero Muñoz, L., Luna Flórez, A., & Vattuone Ramírez, M. Salud de la mujer indígena : intervenciones para reducir la muerte materna. © Banco Interamericano de Desarrollo, 2010

3. Organización Mundial De La Salud. Mortalidad Materna En 2005 : Estimaciones Elaboradas Por La Oms, El Unicef, El Unfpa Y El Banco. Ginebra 27: Ediciones de la OMS, 2008.

4. Ramirez, C. J. Caracterización De Algunas Técnicas Algoritmicas De La Inteligencia Artificial Para El Descubrimiento De Asociaciones Entre Variables Y Su Aplicación En Un Caso De Investigación Específico. Medellin: Tesis Magistral, 2009.

5. Vilca, C. P. Clasificación De Tumores De Mama Usando Métodos. Lima: Tesis, 2009.

6. Dirección Regional De Salud Cusco. Análisis De Situación De La Mortalidad Materna Y Perinatal, Región Cusco, 2007.

7. Ramírez Ramírez, R., & Reyes Moyano, J. Indicadores De Resultado Identificados En Los Programas Estratégicos. Lima: Encuesta Demográfica Y De Salud Familiar – Endes, (2011).

8. Reproductive Health Matters, Mortalidad Y Morbilidad Materna:Gestación Más Segura Para Las Mujeres. Lima: © Reproductive Health Matters, 2009

9. Msp Mynor Gudiel M., 1. E.. Modelo Predictor De Mortalidad Materna. Mortalidad Materna, 22-29, 2001-2002

10. Ministerio De Salud..Dirección General De Epidemiología Situación De Muerte Materna, Perú, 2010 -2011

11. Corso, C. L. Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba, 2009.

12. María N. Moreno García* L. A. Obtención Y Validación De Modelos De Estimación De software Mediante Técnicas De Minería De Datos. Revista colombiana de computación, 3[1], 53-71, 2005

13. Calderón, S. G, Una Metodología Unificada para la Evaluación de Algoritmos de Clasificación tanto Supervisados como No-Supervisados. México D. F.: resumen de tesis doctoral. 2006

14. Calderón Saldaña, J., & Alzamora de los Godos Urcia, L. Regresión Logística Aplicada A La Epidemiología. Revista Salud, Sexualidad y Sociedad, 1[4], 2009

15. Antonio Serrano, E. S., Redes Neuronales Artificiales. Universidad de Valencia, 2010.

16. Calderón, S. G, Una Metodología Unificada para la Evaluación de Algoritmos de Clasificación tanto Supervisados como No-Supervisados. México D. F.: resumen de tesis doctoral. 2006.

17. María García Jiménez, & Aránzazu Álvarez Sierra. Análisis de Datos en WEKA – Pruebas de Selecitividad. Articulo, 2010.

18. Porcada, V. R. Clasificación Supervisada Basada En Redes Bayesianas. Aplicación En Biología Computacional. Madrid: Tesis Doctoral, 2003.

4. SÍNTESIS CURRICULARES DE LA AUTORA:

Pilar Hidalgo León, (16 enero 86'), Ingeniera de Sistemas de la Universidad Andina del Cusco en Perú, Sustentación de proyecto de máster en la Universidad del País Vasco. Contacto: pilarv_hidalgoleon@hotmail.com, pvhidalgo001@ikasle.ehu.es, Urb. LarapaMz. B-5 lote 27, San Jerónimo, Cusco, Perú.