##### Proyecto:

**Pima Indians Diabetes Database**

Cristóbal Felipe Rafael Vásquez Fernández

17.708.318-6

Profesor: Gonzalo Acuña  
Ayudante: Carlos Pérez

Índice de contenido

1. MOTIVACIÓN DEL TEMA 3

2. OBJETIVOS 4

2.1. OBJETIVO GENERAL 4

2.2. OBJETIVOS ESPECÍFICOS 4

3. ANTECEDENTES 5

3.1. VARIABLES DE ENTRADA 5

4. ESTADO DEL ARTE 7

5. ANÁLISIS Y PRE-PROCESAMIENTO DE LOS DATOS 8

6. DISEÑO Y ENTRENAMIENTO DE LA RED NEURONAL 10

7. RESULTADOS 14

7.1. MODELO BASE 14

7.2. MODELO CLASES BALANCEADAS 15

7.3. MODELO CLASES BALANCEADAS CON VALIDACIÓN CRUZADA 16

8. DISCUSIÓN 17

9. CONCLUSIONES 18

10. BIBLIOGRAFÍA 19

11. APÉNDICE 21

11.1. DATOS DE ENTRADA 21

11.1.1. Number of times pregnant 21

11.1.2. Plasma glucose concentration 22

11.1.3. Diastolic blood pressure 22

11.1.4. Triceps skin fold trickness 23

11.1.5. 2-hours serum insulin 23

11.1.6. Body mass index 24

11.1.7. Diabetes pedigree function 24

11.1.8. Age 25

11.2. MEJOR MODELO 25

# Motivación del tema

La habilidad para predecir y discriminar si un sujeto sufre de alguna condición clínica es esencial en situaciones médicas. En este aspecto técnicas estadísticas estándares tales como :análisis discriminante, análisis de regresión, y el análisis factorial se han utilizado para proporcionar esta capacidad. Sin embargo, incluso con la existencia de relaciones funcionales ocultas que pueden proporcionar capacidad predictiva, estas técnicas pueden no tener éxito. Los métodos estadísticos pueden proveer resultados pobres cuando:

* El tamaño de la muestra es pequeño
* La forma de la relación funcional subyacente se desconoce
* Las relaciones funcionales subyacentes implican interacciones complejas e interrelaciones entre un número de variables.

Estas condiciones no son inusuales en problemas médicos. En Tales situaciones médicas, algunos enfoques utilizando redes neuronales ofrecen resultados alentadores (Simmons, 1963).

De esta manera se utilizará un MLP para predecir el comienzo de diabetes mellitus tipo 2 (no insulino dependiente) (Wikipedia, Diabetes mellitus, 2015), en un período de 5 años. Los datos que se utilizan para el estudio son del pueblo Pima y se obtuvieron del repositorio *UCI Machine Learning* (Sigillito, 1990).

El enfoque utilizado consiste principalmente en un balance de clases utilizando validación cruzada, a partir de esto se escogen los mejores modelos generados al cambiar la cantidad de neuronas en la capa oculta y al considerar o no ciertas variables de entrada al modelo. El propósito es mejorar los modelos abordados en el estado del arte.

# Objetivos

## Objetivo general

Modelar eficientemente la predicción de diabetes mellitus no insulino dependiente a un período de 5 años, intentando mejorar los modelos del estado del arte en términos de precisión. Para lo cual se utilizará un Perceptrón Multicapa evaluando el desempeño del modelo a través de la curva ROC y matriz de confusión.

## objetivos específicos

* Realizar un modelo base que no considere un pre-procesamiento de los datos, balance de clases ni validación cruzada.
* Obtener un modelo luego de pre-procesar los datos y realizar balance de clases.
* Generar modelos utilizando balance de clases con validación cruzada.
* Evaluar la precisión de los modelos al considerar y no considerar la variable de entrada “*2-Hours Serum Insulin”*, a través de la curva ROC y matriz de confusión.
* Escoger los mejores modelos para cada caso y realizar un análisis comparativo.
* Evaluar si se mejoran los resultados actuales del estado del arte.

# Antecedentes

La población de estudio son el pueblo Pima ubicado en el estado de Arizona (Estados Unidos) (Pima people, 2016). Este pueblo esta bajo estudio desde el año 1965 por el *Institute National of Diabetes and Digestive and Kidney Diseases,* esto dado que tiene una alta tasa de incidencia de diabetes (Bennet, 1971).

La diabetes fue diagnosticada de acuerdo al criterio de la Organización Mundial de Salud, que corresponde a verificar si luego de una carga de 2 horas de glucosa en el plasma, se presenta una concentración de al menos 200 mg/dl. En caso que la diabetes se manifestara un año después del examen, ese examen fue excluido del estudio, dado que son casos muy sencillos de predecir. De esta manera se consideran los examenes que luego de 5 años revelaron diabetes mellitus o no la revelaron.

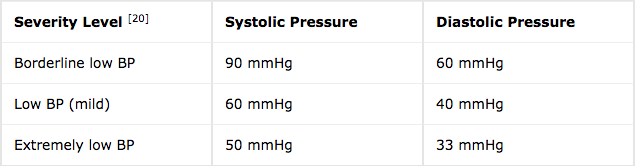
En cuanto a los sujetos de estudio se consideran solo mujeres del pueblo Pima mayores de 21 años.

## variables de entrada

A continuación se realiza un análisis biológico de las variables de entrada, para así tener parámetros de referencia a considerar durante la limpieza de los datos, esto dado que el *data set* que se utiliza consta de *missing values*, sin embargo no están identificados con la notación NaN, por lo que se debe conocer los rangos biológicos para así identificar *outliers* y realizar un tratamiento de estos en el pre-procesamiento.

* *Number of times pregnant:* En cuanto a la cantidad máxima de hijos que puede tener un mujer es difícil de determinar, sin embargo existe casos en que se habla de hasta 87 hijos(Paris, 2010)*.*
* *Plasma glucose concentration at 2 hours in an oral glucose tolerance test:* Según un estudio realizado por (Boshell, 1968) que analiza el suero de insulina en personas obesas y con diabetes mellitus, se tiene que el nivel mínimo para sujetos normales o enfermos no puede ser en ningun caso 0, teniendo como mínimo valores cercanos a 70 mg/dl.
* *Diastolic blood pressure (mm Hg):* De (Pregmed, 2014) se obtiene la Tabla 3.1 que indica que un nivel extremadamente bajo de presión diastólica es de 33 mmHg, mientras que el normal está alrrededor de 120 mmHg.

*Tabla 3.1: Niveles mínimos de presión diastólica y sistólica según severidad.*

**

* *Triceps skin fold trickness (mm):* De (Addo, 2010) se obtiene que los niveles mínimos de esta variable puede rodear valores cercanos a cero, sin embargo nunca llega a ser 0, además se muestra que a medida que aumenta la edad este valor incrementa. Por lo tanto, para la muestra considerada de 21 años o más este valor nunca podría ser muy cercano a cero.
* *2-Hours serum insulin:* De (Boshell, 1968) se obtiene que los niveles mínimos asociados a esta variable en sujetos normales, con diabetes y obesos luego de 2 horas tiene valores mínimos de aproximadamente 50 *uU/ml.* Por otro lado, los valores máximos pueden ascender hasta 1200 *uU/ml.*
* *Body mass index (kg/ m2):* De (Body mass index, 2016) se obtiene que un índice de masa corporal bajo 16.5 kg/m2, indica un nivel de bajo peso severo, mientras que sobre 40 kg/m2 indica obesidad muy severa.
* *Diabetes pedrigee function:* Este índice es desarrollado Smith et al en 1988 considerando las variables de entrada descritas, no tiene un valor máximo asociado y puede tener valores mínimos cercanos a cero.
* *Age:* La edad mínima como criterio de entrada al estudio es de 21 años, mientras que no existe una edad máxima para formar parte de este.

# Estado del arte

Para el estado del arte se consideró principalmente el paper recomendado en el repositorio donde se obtuvo el *data set* (Sigillito, 1990)*.*

En 1988 se abordó el problema de predicción de diabetes mellitus no insulino dependiente considerando 8 variables de entrada con evidencias biológicas que indican un alto nivel de correlación a padecer este enfermedad. En este estudio se considera una población del pueblo Pima del estado de Arizona (Estados Unidos), ya que tiene una alta incidencia de diabetes (Smith, 1988).

Para abordar este problema de clasificación utilizan el algoritmo ADAP, que consiste basicamente en aprender iterativamente del error adaptando los pesos de la red, pero su principal característica está en las entradas excitan sensores que son representados como particiones, estas particiones están conectadas con unidades de asociación las cuales están conectadas con una sola salida. Los unidades de asociación son activas por medio de una función umbral. La preparación de la entrada inicial del algoritmo debe realizar una categorización de las variables, de esta manera se considera por ejemplo 5 categorias de edad que se determinan por rangos de edad. Los resultados que obtienen de sensibilidad y especificidad son de un 76%.

# análisis y Pre-procesamiento de los datos

Como se mencionó anteriormente los datos obtenidos de (Sigillito, 1990), dan una breve descripción de las características del *data set*, en la cual se menciona la presencia de *missing values*, sin embargo no los identifica como NaN, por lo que a partir del análisis biológico de las variables de entrada que se realizó anteriormente se plantea la interpolación de los datos utilizando el método *neirest neighbor*. Cabe destacar que el *data set* contiene 768 sujetos.

Previo a detallar la interpolación de cada variable la Tabla 5.1 muestra un resumen estadístico de las variables de entrada, para así cuantificar la cantidad de *missing values* y varianza de cada variable.

*Tabla 5.1: Resumen estadístico del data set previo al pre-procesamiento de datos.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables de entrada | Promedio | Desviación estandar | Mínimo | Máximo | Cantidad missing values |
| Number of times pregnant | 3.8 | 3.4 | 0 | 17 | 0 |
| Glucose concentration | 120.9 | 32 | 0 | 199 | 5 |
| Diastolic blood pressure | 69.1 | 19.4 | 0 | 122 | 38 |
| Triceps skin fold trickness | 20.5 | 16 | 0 | 99 | 227 |
| 2-Hours Serum insulin | 79.8 | 115.2 | 0 | 846 | 374 |
| Body mass index | 32 | 7.9 | 0 | 67.1 | 11 |
| DPF | 0.5 | 0.3 | 0.078 | 2.42 | 0 |
| Age | 33.2 | 11.8 | 21 | 81 | 0 |

De la Tabla 5.1 se desprende que las variables con mayor cantidad de *missing values* son *Triceps skin fold trickness* y *2-Hours Serum insulin.* Es por esta razón que se decidió realizar modelos que consideraran la variable *2-Hours Serum insulin* y otros que no. A continuación se detalle los criterios de interpolación para cada variable.

* *Number of times pregnant:* Esta variable no fue interpolada dado que, no se identificó *outliers.*
* *Plasma glucose concentration at 2 hours in an oral glucose tolerance test:* Esta variable se interpoló cuando los valores son igual a cero.
* *Diastolic blood pressure (mm Hg):* el valor máximo no se identificó como *outliers* al encontrarse dentro de los rangos normales (120 mmHg), sin embargo se identificó 38 sujetos con presión diastólica menor a 30mmHg lo cual está por debajo del límite estudiado en la literatura, por lo tanto se interpola a partir de este criterio.
* *Triceps skin fold trickness (mm):* Esta variable posee 227 *missing values* lo cual es considerable para una muestra de 768 ejemplos, sin embargo se interpolaron cuando su valor es igual a cero.
* *2-Hours serum insulin:* Esta variable es la que posee más *missing values* (374), motivo que gatilló a realizar modelos que la consideraran y otros que no. Sin embargo, esta variable está altamente asociada a la predicción de diabetes mellitus no insulino dependiente por lo que no pudo eliminarse inmediatamente del estudio, de esta manera se realizó igualmente la interpolación cuando los valores eran iguales a cero. A diferencia de los otros casos, la interpolación consideró un valor aleatorio que podía sumar o restar 25 al valor de la media de la entrada (79.8 +- 25), de esta manera se realiza una especie de submuestreo en la variable.
* *Body mass index (kg/ m2):* En este caso se interpoló los datos cuando los valores eran iguales a cero, ya que al realizar la inspección visual se apreció que solo unos pocos son cero y los demás están dentro de rango normales (>14 kg/m2).
* *Diabetes pedrigee function:* No se presentan *missing values.*
* *Age:* No se presentan *missing values.*

La visualización de las variables de entrada que se realizó para la identificación de *outliers* se encuentra en el Apéndice.

# diseño y entrenamiento de la red neuronal

Para seguir los principios de buen diseño de una red neuronal según (Acuña, 2014), se tiene que el número de parámetros aproximado de la red está dado por la cantidad de ejemplos considerados en el conjunto de entrenamiento dividido por 10. Adicionalmente, a partir de estos principios de buen diseño se considera que una sola capa oculta con un número adecuado de neuronas puede aproximar con un grado de precisión arbitrario cualquier función no lineal (Acuña, 2014), por lo que se utiliza una sola capa oculta en la arquitectura de la red.

En cuanto al tamaño del conjunto de entrenamiento se consideró 70% para entrenamiento y 30% para validación, de esta manera se evita sobreajuste del modelo.

Previo a calcular cantidad de neuronas en la capa oculta se trata el problema de clases balanceadas asociado al *data set.*

Dentro del *data set* que contiene 768 muestras para cada variable, solamente posee 268 sujetos pertecenecientes a la clase enfermos, por lo tanto se presentan un problema de clases desbalanceadas. Para abordar esta situación se presentan tres vertientes básicamente, *under-sampling* que elimina patrones de la clase mayoritaria*, over-sampling* que replica patrones en la clase minoritaria y por último, predisponer internamente el proceso de descriminación para compensar el desequilibrio (Rosas, 2006). El enfoque se adoptó es el de *under-sampling*, por lo que se eliminó patrones en la clase mayoritaria reduciendo la muestra a 536 (268 muestras por clase), donde 375 se utilizan para entrenamiento y 161 para validación de acuerdo a los porcentajes que se mencionó anteriormente.

Con el problema de clases desbalanceadas abordado se pasa a calcular la cantidad mínima de neuronas en la capa oculta utilizando la siguiente ecuación (Acuña, 2014):

Donde:

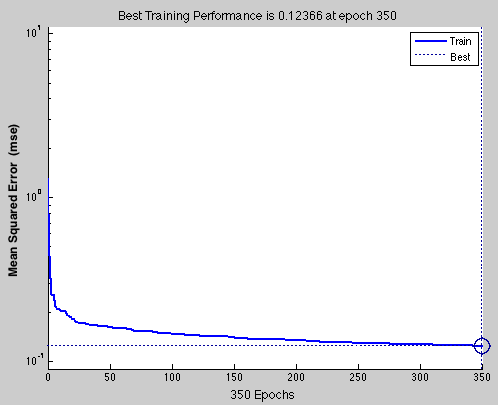
* : Cantidad de parámetros de la red.
* : Cantidad de entradas de la red
* : Cantidad de salidas de la red

Al reemplazar los valores en la fórmula se tiene lo siguiente:

Por lo tanto, la cantidad de neuronas que se debiesen considerar en la capa oculta inicialmente son 3, sin embargo esto es impracticable, dado que la cantidad de entradas es de 8 y se recomienda utilizar un número mínimo de neuronas en la capa oculta equivalente a la cantidad de entradas de la red. Este problema surge por la falta de ejemplos, por lo que para hacer frente a esta situación se utilizó el método de validación cruzada dividiendo el conjunto de datos en 10 muestras, conservando el tamaño del conjunto de entrenamiento y validación. De esta manera se definió 8 neuronas en la capa oculta dentro de la validación cruzada.

En cuanto al entrenamiento de la red, se crea una red pre-alimentada utilizando el comando *newff*. Para actualizar los pesos y *bias* de la red se utiliza la función de entrenamiento *trainscg* que aplica el algoritmo de *scaled conjugate gradient ,* ya que según los *benchmark* realizados por(MathLab, 2015)este algoritmo tiene buenos resultados al aplicarse sobre un *data set* de diabetes, que considera variables similares al modelo que se plantea. Cabe destacar que si bien, el algoritmo más eficiente para el *data set* de diabetes fue el de *resilient-backpropagation,* dio resultados pobres al aplicarse en el modelo por lo que fue desechado. Además se utilizó *learngd* como función de aprendizaje para el gradiente descendente. Por último, para medir la performance de la red, se utiliza el error cuadrático medio a través de la función *mse*.

Luego de tener la red creada, se procede a entrenarla utilizando el conjunto de entrenamiento seleccionado con 350 iteraciones. A partir de este entrenamiento, luego de realizar 10 pruebas se llega a la conclusión que cuando el error se encuentra alrededor de 10-1 se mantiene prácticamente constante, por lo que es innecesario aumentar el número de iteraciones para que la red intente seguir disminuyendo el error. En la Figura 6.1 puede observarse que el error tiende a mantenerse constante alrededor de la iteración 350.



*Figura 6.1: MSE en medida que aumentan las iteraciones.*

En base al diseño de la red descrito anteriormente, se generaron los siguientes modelos:

1. Modelo base: modelo realizado en primera instancia sin pre-procesar datos ni balancear clases.
2. Modelo clases balanceadas: modelo que considera pre-procesamiento de datos y balance de clases.
3. Modelo clases balanceadas con validación cruzada: modelo que considera pre-procesamiento de datos, balance de clases y validación cruzada.

Para cada modelo se realizó tres pruebas y se escogió el mejor de las tres, esto se hizo ya que los pesos iniciales de la red son aleatorios. Para los modelos 2 y 3 se aumenta la cantidad de neuronas en la capa oculta para evaluar si existe una mejora en el rendimiento. Finalmente para cada modelo se evaluó el rendimiento al incluir y no incluir la variable *2-hours serum insulin*.

# Resultados

Para sintetizar la muestra de resultados se muestra solamente tablas de resumen. En el apéndice se encuentra la matriz de confusión y curva ROC del mejor modelo.

En esta sección se abreviará el uso de la variable 2-hours serum insulin a: I (solamente el nombre de la variable), CI (considerando la variable en el modelo) y SI (sin considerar la variable en el modelo). De esta manera se referirá a los modelos como modelo CI o modelo SI.

## modelo base

En la Tabla 7.1 se muestran los resultados obtenidos para el modelo base.

*Tabla 7.1: Resumen resultados Modelo Base*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelos Base | | | | | |
|  | **Precisión** | **Sensibilidad** | **Especificidad** | **PPV** | **NPV** |
| CI | 80.1% | 91.7% | 55.4% | 80.1% | 77.6% |
| SI | 80.5% | 88.7% | 52.1% | 78.4% | 69.6% |

Como se puede apreciar en la tabla, hay un alto nivel de precisión para ambos modelos al CI y SI, sin embargo estos modelos tienden a incurrir en falsos positivos lo cual se ve reflejado en su bajo nivel de especificidad, a pesar de esto es un modelo base aceptable teniendo en cuenta que no existe un balance de clases, pre-procesamiento de los datos ni validación cruzada. Puede decirse que este modelo puede ser aplicado cuando se desea altos niveles de sensibilidad y bajos de especificidad, es decir cuando se desea identificar efectivamente como enfermas a personas diabéticas, por otro lado no es bueno para detectar ausencia de enfermedad en sujetos sanos

En cuanto a la influencia de la variable I en el modelo, se puede apreciar que los resultados para ambos modelos son similares por lo que puede no estar aportando información relevante a la predicción de la diabetes, sin embargo al no considerarse una limpieza de los datos en estos modelos no puede darse un veredicto al respecto.

## modelo clases balanceadas

Al realizar el balance de clases y considerar 8 neuronas en la capa oculta se obtienen los resultados que se muestran en la tabla 7.2.

*Tabla 7.2: Resumen resultados Modelo Clases Balanceadas con 8 neuronas en la capa oculta*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelos Clases Balanceadas NC=8 | | | | | |
|  | **Precisión** | **Sensibilidad** | **Especificidad** | **PPV** | **NPV** |
| CI | 79.0% | 82.9% | 75.0% | 77.3% | 81.1% |
| SI | 82.7% | 78.9% | 88.1% | 90.4% | 74.7% |

Al considerar 16 neuronas en la capa oculta se obtienen los resultados que se muestran en la tabla 7.3.

*Tabla 7.3: Resumen resultados Modelo Clases Balanceadas con 16 neuronas en la capa oculta*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelos Clases Balanceadas NC=16 | | | | | |
|  | **Precisión** | **Sensibilidad** | **Especificidad** | **PPV** | **NPV** |
| CI | 77.2% | 77.8% | 76.5% | 76.8% | 77.5% |
| SI | 75.3% | 75.3% | 75.3% | 73.4% | 77.1% |

Como se puede apreciar al aumentar el número de neuronas en la capa oculta la precisión de los modelos puede disminuir hasta en un ~7%. Además se refleja una mejora considerable del modelo al realizar un balance de clases y limpieza de los datos (interpolación), lo cual se ve reflejado principalmente en la especificidad. El mejor modelo para este caso corresponde al de 8 neuronas en la capa oculta SI, que supera en casi todo aspecto al modelo base excepto en la sensibilidad. Aún sigue la interrogante de si la variable I da verdadero aporte al modelo, dado que en este caso el mejor modelo no la considera, por lo que la alta cantidad de *missing values* puede reflejarse en un bajo aporte al modelo, a pesar de la alta correlación biológica de la variable con la enfermedad.

## Modelo clases balanceadas con validación cruzada

Finalmente se presentan los resultados al realizar: limpieza de los datos, balance de clases y validación cruzada, con 8 y 16 neuronas en la capa oculta. La Tabla 7.4 y 7.5 muestran estos modelos respectivamente.

*Tabla 7.4: Resumen resultados Modelo Clases Balanceadas con validación cruzada considerando 8 neuronas en la capa oculta.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelos Clases Balanceadas con validación cruzada NC=8 | | | | | |
|  | **Precisión** | **Sensibilidad** | **Especificidad** | **PPV** | **NPV** |
| CI | 85.2% | 77.6% | 91.9% | 79.2% | 82.3% |
| SI | 83.2% | 87.0% | 79.8% | 79.8% | 87.0% |

*Tabla 7.5: Resumen resultados Modelo Clases Balanceadas con validación cruzada considerando 16 neuronas en la capa oculta.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelos Clases Balanceadas con validación cruzada NC=16 | | | | | |
|  | **Precisión** | **Sensibilidad** | **Especificidad** | **PPV** | **NPV** |
| CI | 80.7% | 87.5% | 74.1% | 76.9% | 85.7% |
| SI | 80.1% | 78.2% | 81.9% | 80.3% | 80.0% |

Los mejores modelos para este caso también son los que consideran 16 neuronas en la capa oculta en el diseño de la red. A diferencia de los modelos anteriores, en este caso es más difícil escoger el mejor modelo, esto ya que si bien el modelo CI tiene mayor precisión y especificidad se tiene menos sensibilidad con respecto al modelo SI, sin embargo se escoge el modelo CI para efectos prácticos de la comparación y conclusiones.

Por otro lado, no se puede aseverar nada respecto a la variable I en cuanto a su valor dentro del modelo, sin embargo se considera que con una menor cantidad de *missing values* el poder predectivo del modelo podría mejorar. De todos los modelos realizados los que presentan mejores resultados son los de esta categoria, y además utilizan sólo 8 neuronas en la capa oculta, lo cual simplifica el modelo y la cantidad de parámetros que calcula.

# discusión

Se considera que existe una falencia en la calidad de los datos, ya que no logró determinarse si la variable *2-hours serum insulin* al tener aproximadamente 48% de *missing values* lograba dar verdadero valor al modelo. Sin embargo, se sabe que biológicamente está altamente correlacionada con la enfermedad de diabetes mellitus no insulino dependiente, por lo que una mejor calidad los datos podría aumentar la capacidad predictiva del modelo.

A pesar de que la cantidad de datos usados para el entrenamiento de la red es pequeña, considerando que se desea tener una cantidad de neuronas en la capa oculta mayor o igual a la cantidad de entradas según la fórmula utilizada de (Acuña, 2014), se obtuvo buenos resultados en los modelos. En este aspecto la validación cruzada ayudó a mejorar significativamente el rendimiento del modelo.

Se logró mejorar los resultados obtenidos por Smith et al. en un 1% respecto a la sensibilidad y un 17% en términos de especificidad, por lo que el enfoque adoptado del MLP utilizando el algoritmo *scaled conjugate gradient* para el entrenamiento de la red, balance de clases *under-sampling* yvalidación cruzada con 10 iteraciones es superior al modelo propuesto por (Jack W. Smith, 1988). Sin embargo, aún se debe validar el modelo con otro *data set*, e incluso considerar otros *data set* en los cuales se incluya hombres y personas de otros sectores geográficos en el estudio.

Aún queda inspeccionar otras variantes para abordar el problema de desbalance de clase dado que, al eliminar ejemplos de manera aleatoria con *under-sampling* se puede estar perdiendo información importante del problema, por lo que se debería comparar resultados al utilizar el enfoque *over-sampling.*

# conclusiones

De (wikipedia, 2015) se obtiene que el predictor modelo es considerado como un test bueno. Adicionalmente se logró el objetivo de modelar eficientemente un predictor de diabetes no insulino dependiente a un periodo de 5 años, superando el modelo de Smith et al. Significativamente en términos de especificidad y precisión.

Se logró apreciar las mejoras del modelo al incluir balance de clases y luego validación cruzada, logrando disminuir signitificativamente la cantidad de falsos positivos que se obtenía en el modelo base.

Los modelos más simple que consideran 8 neuronas en la capa oculta tendieron a generar mejores resultados que los más complejos que consideraban 16.

Se comparo efectivamente bajo que características del modelo se obtenian mejores resultados, resultando de esta manera que el modelo con mejores resultados fue en el que se realizó la mayor cantidad de tratamiento sobre los datos y se abordo los problemas de desbalance de clase con *under-sampling* y falta de ejemplos en la muestra por medio de validación cruzada.

A pesar de los buenos resultados que se obtuvo se encontró falencias en el *data set* principalmente asociadas a la entrada *2-hours serum insulin* que posee una gran cantidad de *missing values* cercana al 48%, sin embargo el mejor modelo que se generó considera esta variable en el modelo a pesar de este problema. Se considera adicionalmente que el método de interpolación utilizado de sumar o restar un valor aleatorio con rango entre +- 25 a la media del valor de entrada, simulando una especie de submuestreo, parece tener resultados satisfactorios en el pre-procesamiento de los datos.

# Bibliografía

Wikipedia. (14 de enero de 2016). *Body mass index*. Recuperado el 16 de enero de 2016, de wikipedia: https://en.wikipedia.org/wiki/Body\_mass\_index

Wikipedia. (23 de diciembre de 2015). *Diabetes mellitus*. Recuperado el 15 de enero de 2016, de Wikipedia: https://es.wikipedia.org/wiki/Diabetes\_mellitus

Wikipedia. (2 de enero de 2016). *Pima people*. Recuperado el 15 de enero de 2015, de wikipedia: https://en.wikipedia.org/wiki/Pima\_people

Yaw Addo, J. H. (diciember de 2010). Reference curves for triceps and subscapular skinfold thicknesses in US children and adolescents. *The American Journal of Nutrition* .

Acuña, G. (2014). *Curso de Ingeniería Neuronal Clase2: Sistemas estáticos.* Univerisdad de Santiago de Chile, Departamento de Ingeniería Informática, Santiago.

Boris.R Boshell, H. B. (1968). Serum insulin in obesity and diabetes mellitus. *The American journal of Clinical Nutrition , 21*, 14919-1428.

Jack W. Smith, J. E. (1988). *Using the ADAP Learning algorithm to forecast the onset of diabetes mellitus.* The Johns Hopkins University School of Medicinet , Maryland.

MathLab. (2015). *Choose a Multilayer Neural Network Training Function*. Recuperado el 13 de enero de 2015, de MathWorks: http://www.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html#bss4gz0-29

P.H Bennet, T. B. (1971). Diabetes mellitus in America (Pima) Indians. *Lancer* *, 2*, 125-128.

Paris, E. (24 de marzo de 2010). *La madre con más hijos de la historia*. Recuperado el 15 de enero de 2016, de Bebés y más: http://www.bebesymas.com/otros/la-madre-con-mas-hijos-de-la-historia

Pregmed. (12 de agosto de 2014). *Pregmed*. Obtenido de Low Blood Pressure during Pregnancy: http://www.pregmed.org/low-blood-pressure-during-pregnancy.htm

Sigillito, V. (9 de mayo de 1990). *Pima Indians Diabetes Data Set* . Recuperado el 11 de enero de 2016, de UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Simmons, G. (1963). *A Mathematical Model for an Associative Memory* (Vols. SRC-641). Sandia Corporation Monograph.

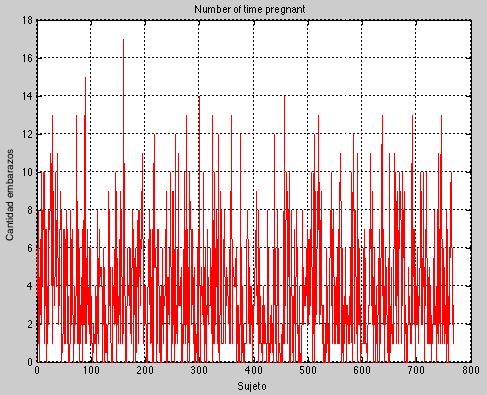
Rosas, R. M. (2006). *Técnicas de submuestreo, toma de decisiones y análisis de diversidad en aprendizaje supervisado con sistemas múltiples de clasificación.* Universitat Jaume, Departamento de Lenguajes y Sistemas Informáticos, Castello de la plana.

# apéndice

En esta sección se incluye la gráfica ROC y matriz de confusión asociada al mejor modelo. Además se encuentra la distribución de los datos de entrada sin la limpieza.

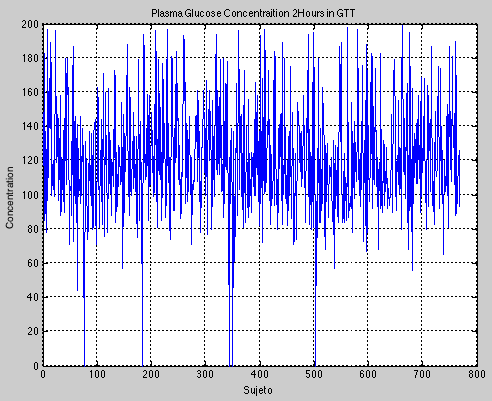
## datos de entrada

### Number of times pregnant



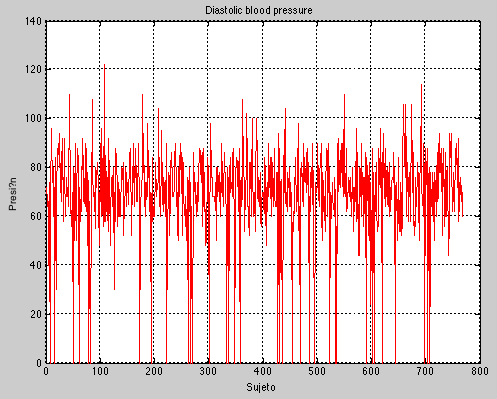
*Figura 11.1: Distrubución de los datos en entrada Number of times pregnant.*

### Plasma glucose concentration



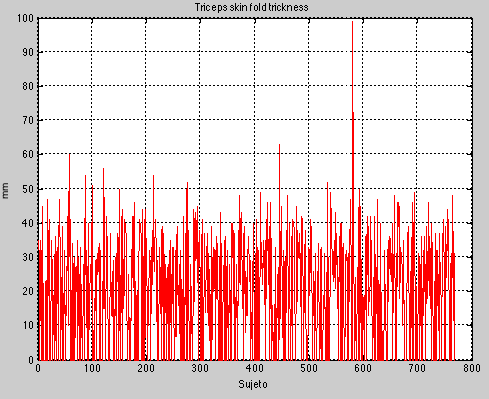
*Figura 11.2: Distrubución de los datos en entrada Plasma glucose concentration.*

### Diastolic blood pressure



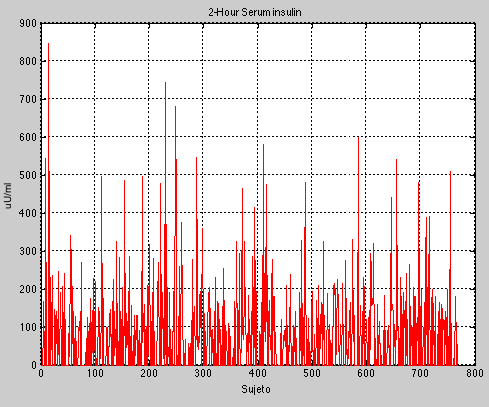
*Figura 11.3: Distrubución de los datos en entrada diastolic blood pressure.*

### Triceps skin fold trickness



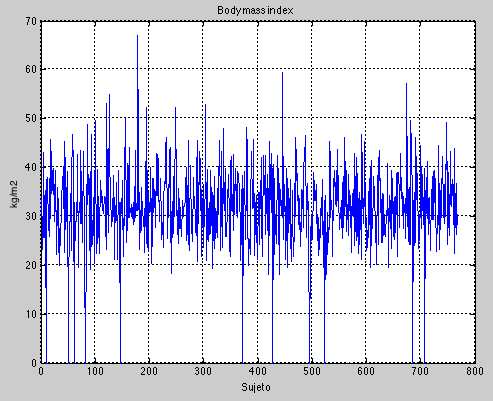
*Figura 11.4: Distrubución de los datos en entrada triceps skin fold trickness.*

### 2-hours serum insulin



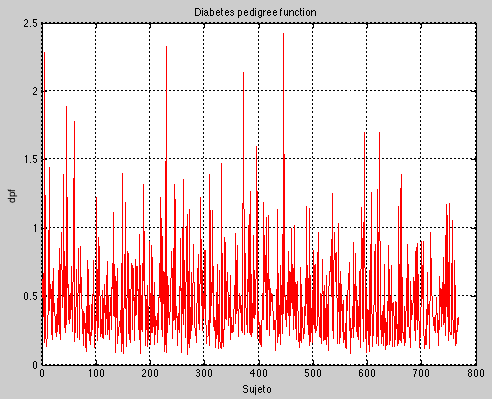
*Figura 11.5: Distrubución de los datos en entrada 2-hours serum insulin.*

### Body mass index



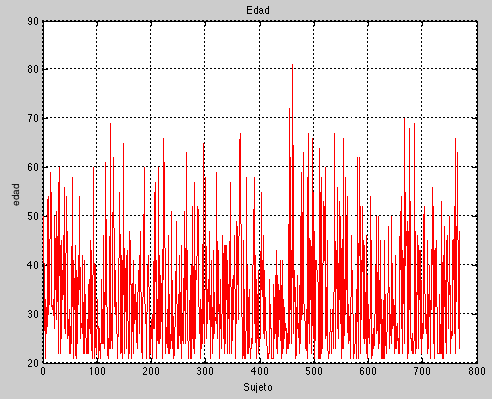
*Figura 11.6: Distrubución de los datos en entrada Body mass index.*

### Diabetes pedigree function



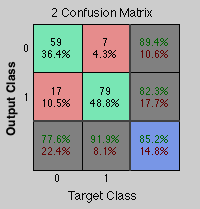
*Figura 11.7: Distrubución de los datos en entrada diabetes pedigree function*

### Age

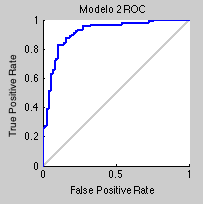


*Figura 11.8: Distrubución de los datos en entrada age*

## mejor modelo



*Figura 11.9: Matriz confusión Modelo clases balanceadas con validación cruzada CI.*

**

*Figura 11.10: Curva ROC Modelo clases balanceadas con validación cruzada CI.*