Uleam
UNIVERSIDAD LAICA
ELOY ALFARO DE MANABÍ

ESPAMMFL
ESCUELA SUPERIOR POLITÉCNICA
AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ

FACCI
FACULTAD DE CIENCIAS INFORMÁTICAS

Carrera de
COMPUTACIÓN

CIDE
Centro de Investigación y Desarrollo Ecuador

EP ULEAM
EMPRESA PÚBLICA DE PRODUCCIÓN Y DESARROLLO ESTRATÉGICO

Mantahost
Hotel

II Congreso Internacional de Tecnologías de la Información y Computación CITIC 2018

CITIC

29, 30 y 31 octubre 2018

Manta – Manabí
Sede:ULEAM

Estructuras de Minería de Datos como soporte para la gestión de un sistema de comercialización de energía eléctrica. Propuesta alternativa.

Ing. Jorge Iván Pincay Ponce, MSc.

Docente ocasional de la Universidad Laica Eloy Alfaro de Manabí, desde el 2008

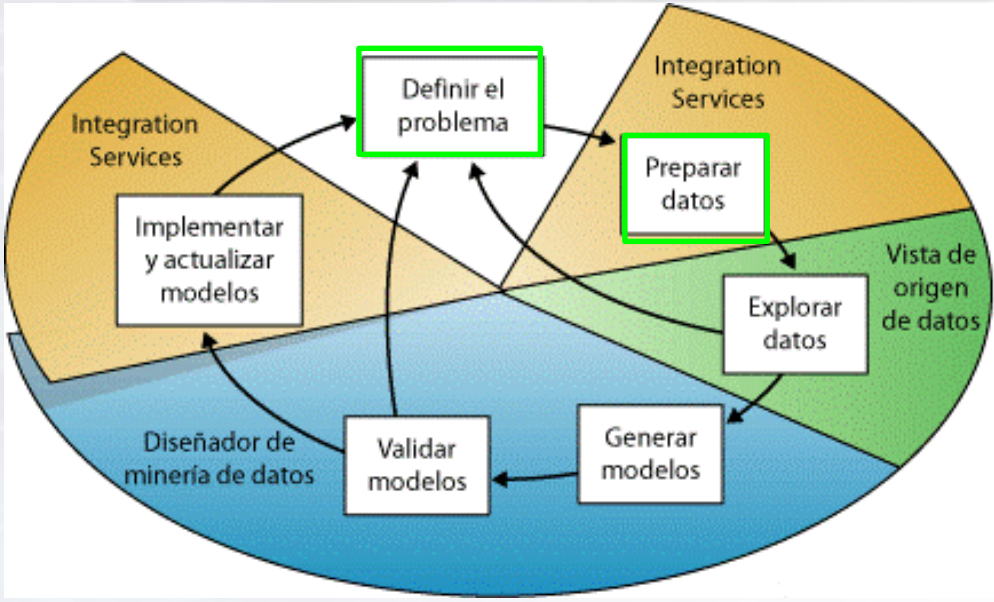
- *Ingeniero en Sistemas por la Universidad Laica Eloy Alfaro de Manabí (Ecuador)*
- *Estudiante certificado en MOOCs de universidades de USA, Bélgica y España.*
- *Diplomado en Educación Universitaria por Competencias por la Universidad del Azuay (Ecuador)*
- *Máster en Gestión de TICs por la Universidad Nacional de Piura (Perú),*
- *Máster en Ingeniería de Software por la Universidad de Alcalá (España),*
- *Doctorando en Informática por Universidad Nacional de Plata (Argentina).*

OBJETIVO DE ESTUDIO

Presentar una propuesta alternativa que sirva como soporte a la gestión del sistema de comercialización de la energía eléctrica en la empresa pública de la ciudad de Manta, a partir de una muestra de datos extraídos de las facturas de consumo residencial correspondientes al año 2015, sobre la cual se aplicaran estructuras de Redes Neuronales Artificiales y de Reglas de Asociación .

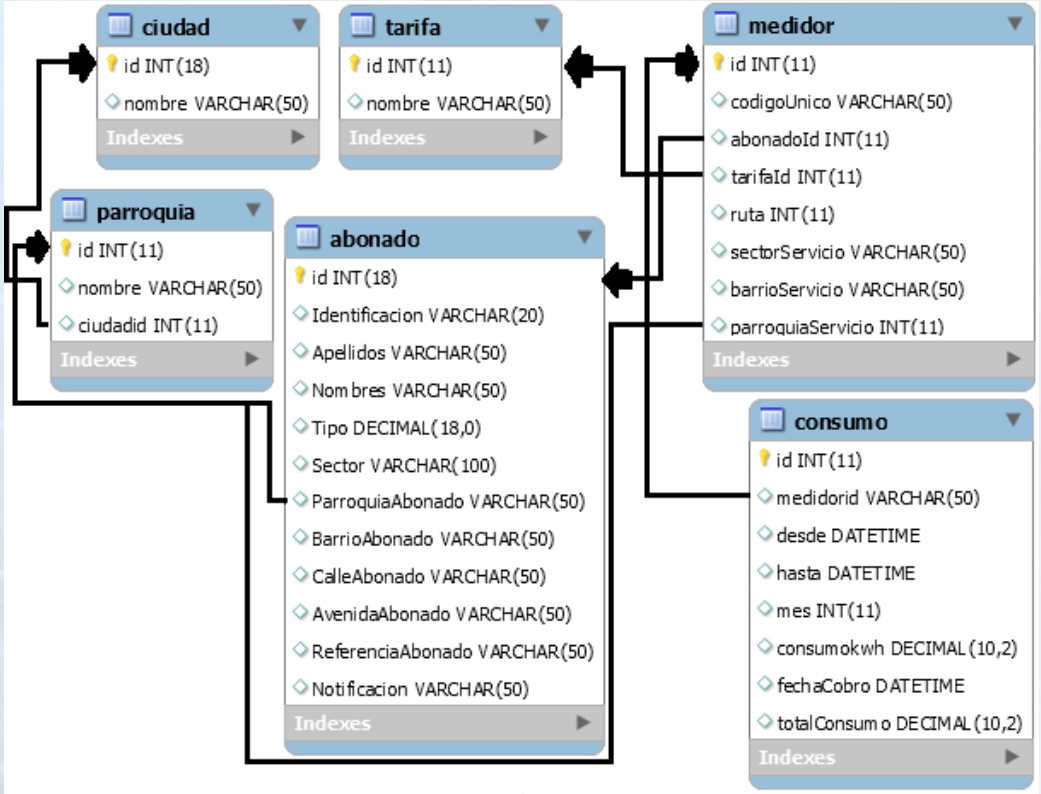
P1. Definir el problema. - Predecir y clasificar mediante redes neuronales y reglas de asociación: Día del Mes, día de la semana, mes del año y estación climática en que más se generan cobros

METODOLOGÍA DE LA INVESTIGACIÓN



Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

P2. Preparar los datos. -




```
SELECT  cast(consumo.id as decimal) as ConsumoID,
                cast(consumo.medidorid as decimal) as MedidorID,
                cast(consumo.desde as nchar(20)) as Desde,
                cast(consumo.hasta as nchar(20)) as Hasta,
                cast(consumo.mes as NCHAR(20)) as Mes,
                cast(consumo.consumokwh as decimal) as kWh,
                cast(consumo.fechaCobro as nchar(20)) as Cobro,
                cast(day(consumo.fechaCobro) as nchar(20)) as DiaDeMes,
                cast(dayofweek(consumo.fechaCobro) as nchar(10)) as DiaDeSemana,
                cast(monthname(consumo.fechaCobro) as nchar(20)) as MesdeAnio,
                CASE mes
                WHEN 1 THEN 'INVIERNO'
                WHEN 2 THEN 'INVIERNO'
                WHEN 3 THEN 'INVIERNO'
                WHEN 4 THEN 'INVIERNO'
                WHEN 5 THEN 'INVIERNO'
                ELSE 'VERANO'
                END as Estacion,
                cast(consumo.totalConsumo as decimal) as Total
FROM consumo
```

P3. Explorar los datos. -

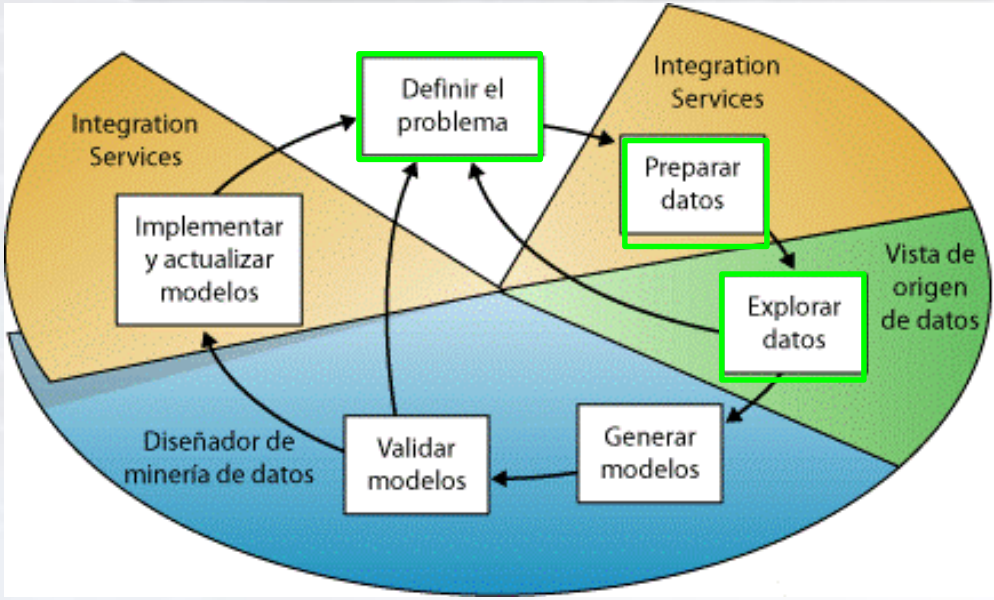
ARFF-Viewer - C:\Users\Jorge Iván\Desktop\cnelepfullEstacionesMeses.arff

File Edit View

cnelepfullEstacionesMeses.arff

Relation: QueryResult

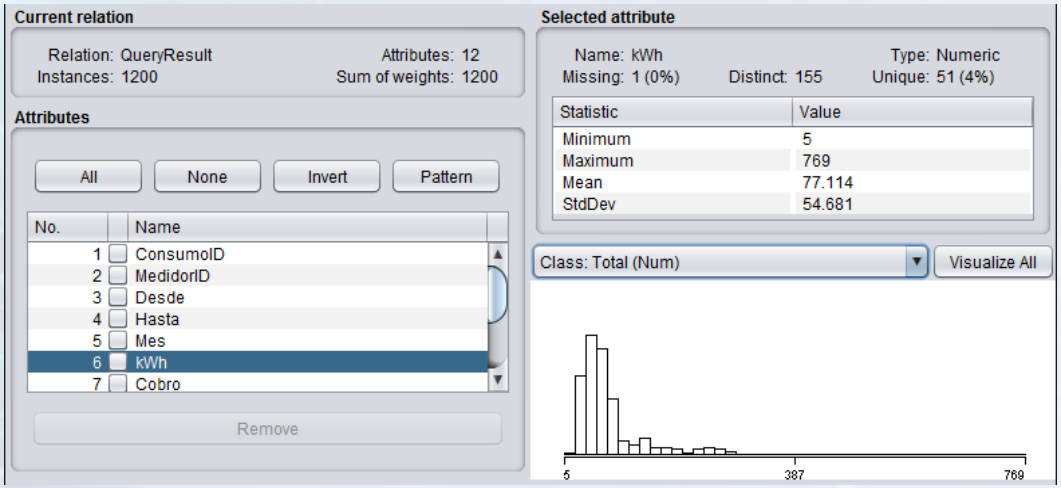
No.	1: ConsumoID Numeric	2: MedidorID Numeric	3: Desde Nominal	4: Hasta Nominal	5: Mes Nominal	6: kWh Numeric	7: Cobro Nominal	8: DiaDel Nomina
1	1.0	1.0	2015-01-01 00:00:00	2015-01-26 00:00:00	1	46.0	2015-12-26 00:00:00	26
2	2.0	1.0	2015-02-01 00:00:00	2015-02-25 00:00:00	2	65.0	2015-02-22 00:00:00	22
3	3.0	1.0	2015-03-01 00:00:00	2015-03-24 00:00:00	3	80.0	2015-03-22 00:00:00	22
4	4.0	1.0	2015-04-01 00:00:00	2015-04-23 00:00:00	4	79.0	2015-04-22 00:00:00	22
5	5.0	1.0	2015-05-01 00:00:00	2015-05-25 00:00:00	5	76.0	2015-05-22 00:00:00	22
6	6.0	1.0	2015-06-01 00:00:00	2015-06-23 00:00:00	6	68.0	2015-06-22 00:00:00	22
7	7.0	1.0	2015-07-01 00:00:00	2015-07-26 00:00:00	7	90.0	2015-07-22 00:00:00	22

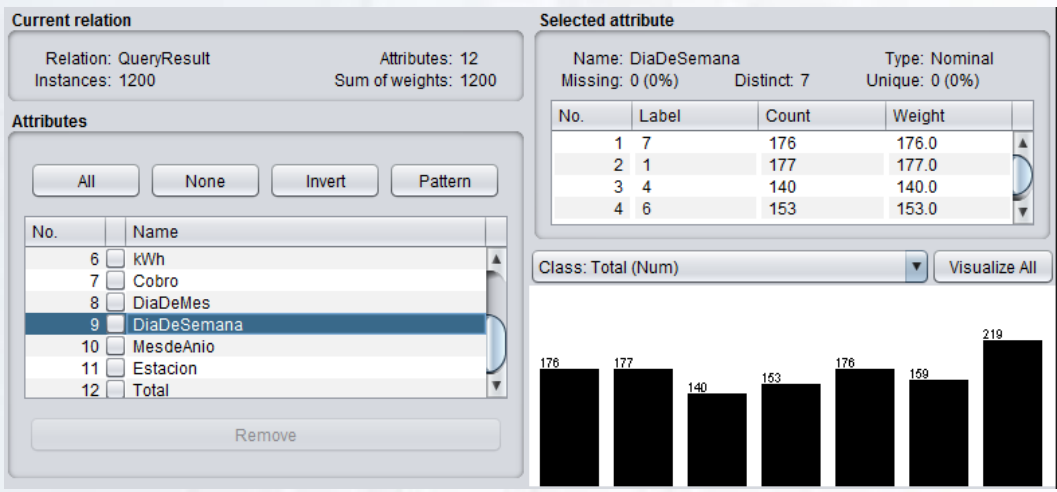


Procesos para la generación de un modelo de MD

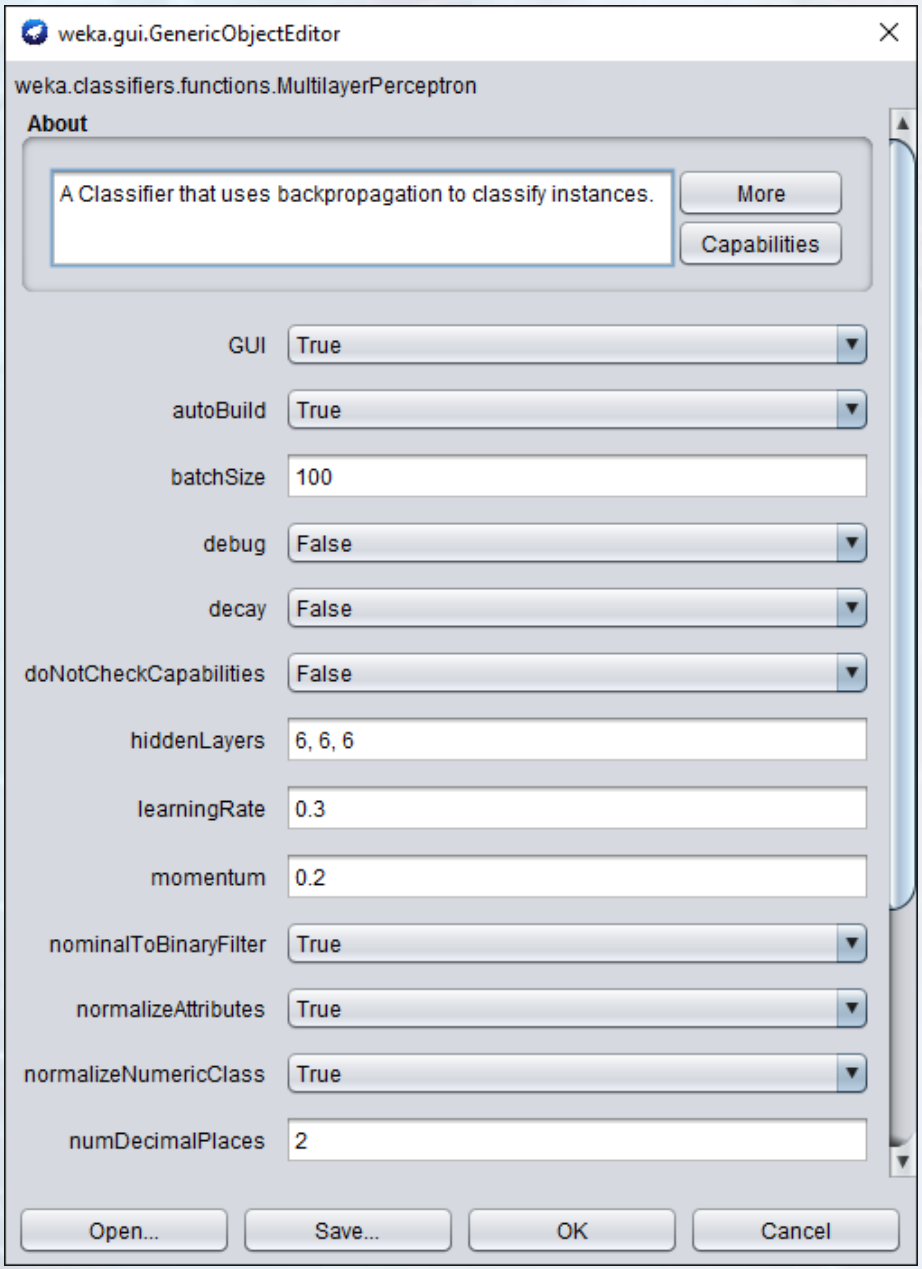
Fuente: https://tinyurl.com/y9s6tys6

Estadísticas del atributo *kW* consumido en cada periodo, relevante para las predicciones y clasificaciones pese a la cercanía entre la desviación estándar y el promedio, pues un 49% tienen por lo menos dos registros, lo que significa que el 49% de los abonados tiene al menos un consumo similar con otro abonado.

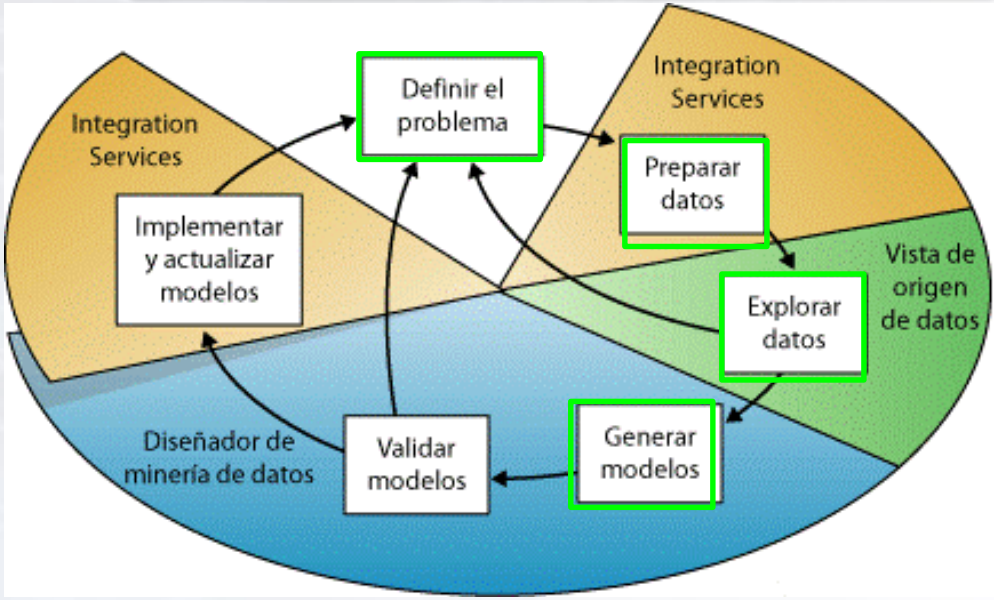




P4. 1 Generar Modelos. -

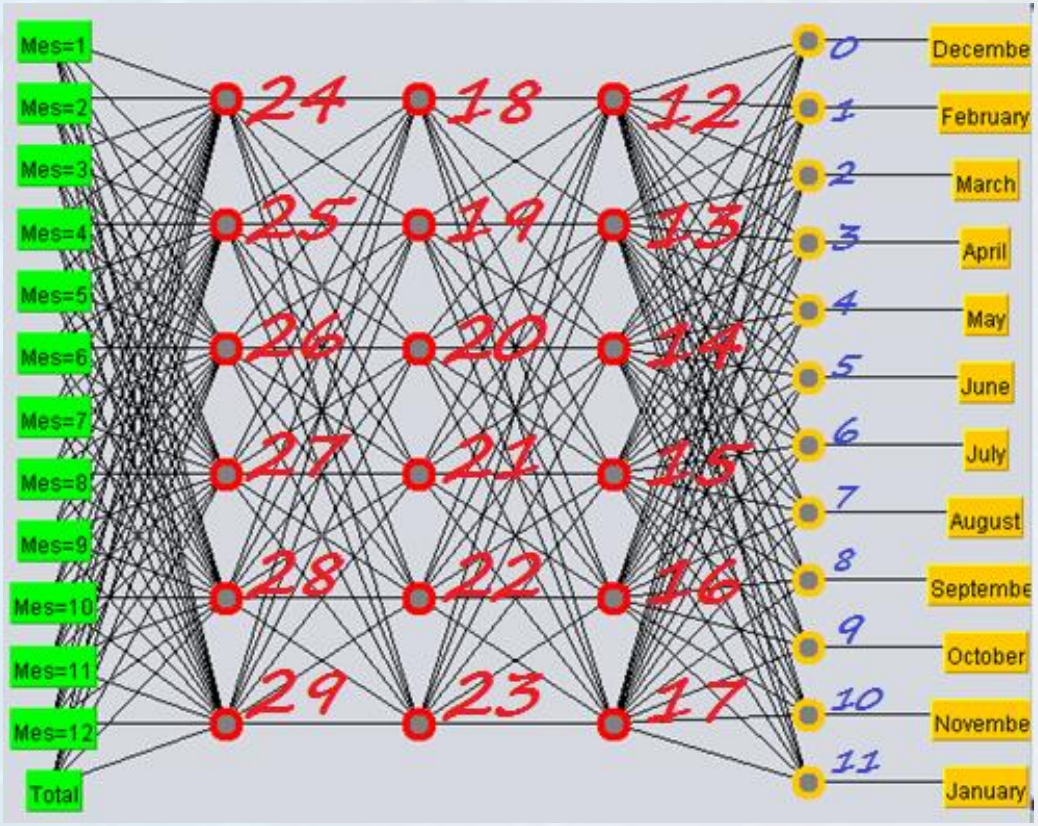


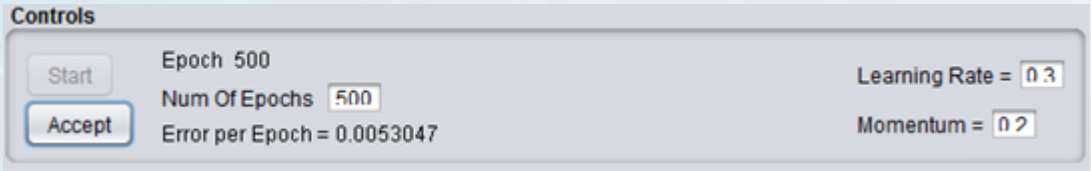
Red Neuronal de clase *Perceptron Multicapa*, que usa *backprogation* para clasificar las 1200 instancias



Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

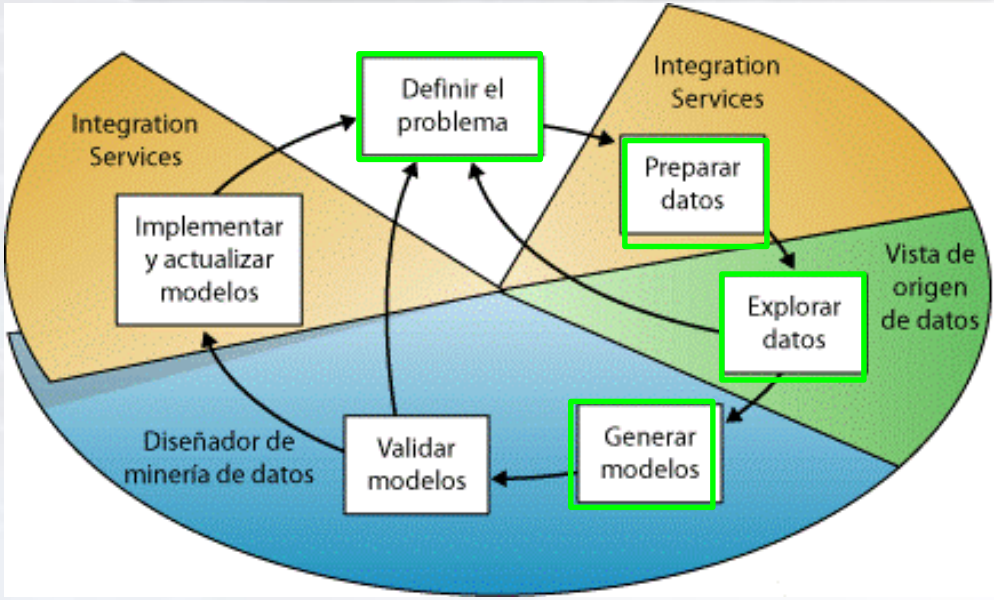
P4. 2 Generar Modelos. -





A modo de ilustración, WEKA identifica los 29 nodos, donde 12, es decir del 0 al 11 son clases de salida y 18 se corresponden con las tres capas ocultas de seis neuronas cada una. El orden sucede porque se aplica el algoritmo de *backpropagation*

METODOLOGÍA DE LA INVESTIGACIÓN

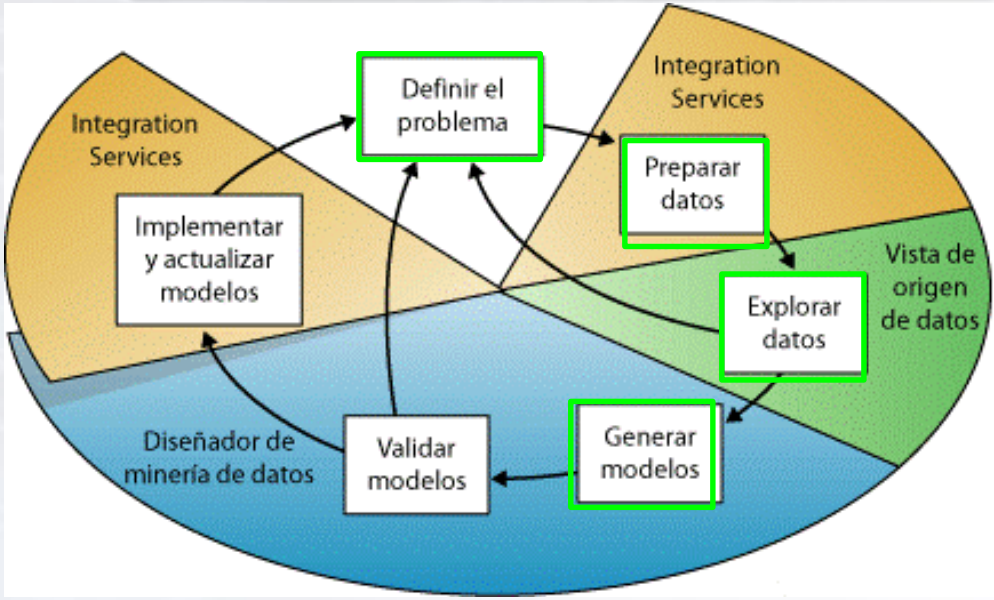


Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

P4. 3 Generar Modelos. -

El segundo modelo generado fue el de Reglas de Asociación con el algoritmo PART, empleado para determinar los días de la semana en que más se cobra el servicio eléctrico. Sus configuraciones fueron:

- Número de instancias: 1200.
- Atributos de entradas: Mes a pagar, kW consumidos, día del mes del cobro, mes del año del cobro, la estación climática.
- Atributos de salida: día de la semana del cobro.
- Entrenamiento: Use training set.
- Algoritmo: PART.
- Número de épocas: 500, lo que significa que los 1200 registro se introducen 500 veces hasta procurar que el error cuadrático medio sea lo menor posible.



Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

```
Mes = 9 AND
DiaDeMes <= 22 AND
DiaDeMes > 20 AND
DiaDeMes > 21: 3 (11.0)
```

Cuando tocaba pagar septiembre, los abonados que pudieron hacerlo entre los días 20 y 22 (11 en total) prefirieron pagar un miércoles

```
Mes = 5 AND
DiaDeMes > 24 AND
DiaDeMes <= 26 AND
DiaDeMes > 25: 3 (30.0)
```

Cuando tocaba pagar mayo, los abonados que pudieron hacerlo entre los días 24 y 26 (30 en total) prefirieron pagar un miércoles

```
Estacion = INVIERNO AND
DiaDeMes <= 24 AND
DiaDeMes > 22 AND
DiaDeMes > 23: 7 (12.0)
```

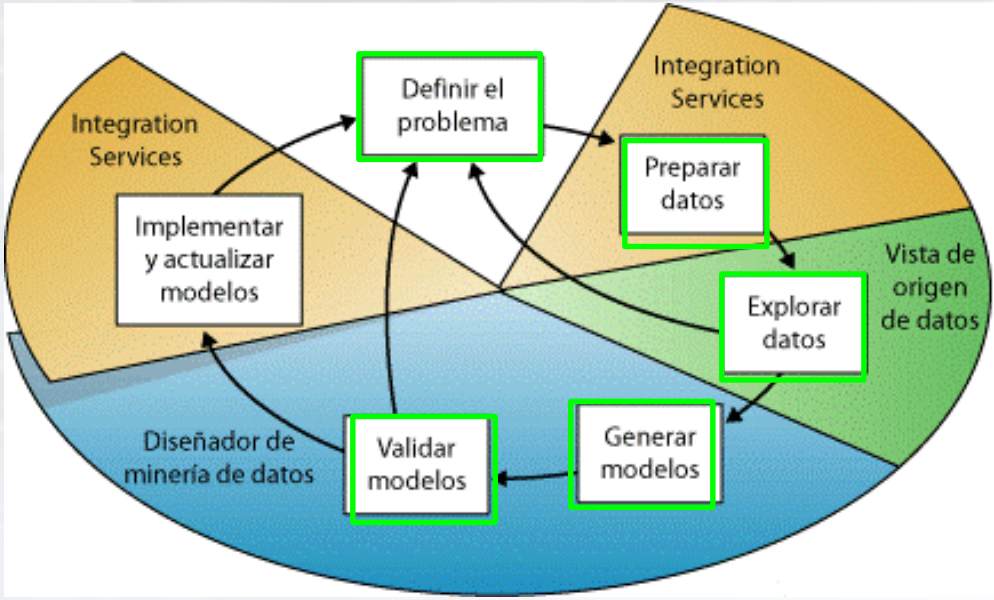
Cuando tocaba pagar algún mes del invierno, los abonados que pudieron hacerlo entre los días 22 y 24 (12 en total) prefirieron pagar un domingo.

P5. Validar Modelos. -

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   <-- classified as
 100   0   0   0   0   0   0   0   0   0   0   1 |   a = December
   0  99   0   0   0   0   0   0   0   0   0   0 |   b = February
   0   0 100   0   0   0   0   0   0   0   0   0 |   c = March
   0   0   0  99   0   0   0   0   0   0   0   0 |   d = April
   0   0   0   1 100   0   0   0   0   0   0   0 |   e = May
   0   0   0   0   0 100   0   0   0   0   0   0 |   f = June
   0   0   0   0   0   0 100   0   0   0   0   0 |   g = July
   0   0   0   0   0   0   0 100   0   0   0   0 |   h = August
   0   0   0   0   0   0   0   0 100   0   0   0 |   i = September
   0   0   0   0   0   0   0   0   0 100   1   0 |   j = October
   0   0   0   0   0   0   0   0   0   0  99   0 |   k = November
   0   1   0   0   0   0   0   0   0   0   0  99 |   l = January
```

Matriz de confusión correspondiente al Perceptron Multicapa, que muestra la aceptable clasificación de los datos, por ejemplo, en mayo (fila) se registraron 101 valores cobrados, de los cuales el modelo ha clasificado correctamente como e (e=mayo) a 100 e incorrectamente clasificó 1 caso como d (d = abril). En noviembre no hay errores de clasificación.

METODOLOGÍA DE LA INVESTIGACIÓN



Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

Resumen de los 1200 registros analizados con el *Perceptron Multicapa*. El *root mean squared error* es del 0,0637 en tanto que el error absoluto es muy pequeño pero positivo.

```
=== Summary ===

Correctly Classified Instances        1196               99.6667 %
Incorrectly Classified Instances         4                0.3333 %
Kappa statistic                          0.9964
Mean absolute error                      0.0192
Root mean squared error                  0.0637
Relative absolute error                 12.5856 %
Root relative squared error             23.0315 %
Total Number of Instances             1200
```

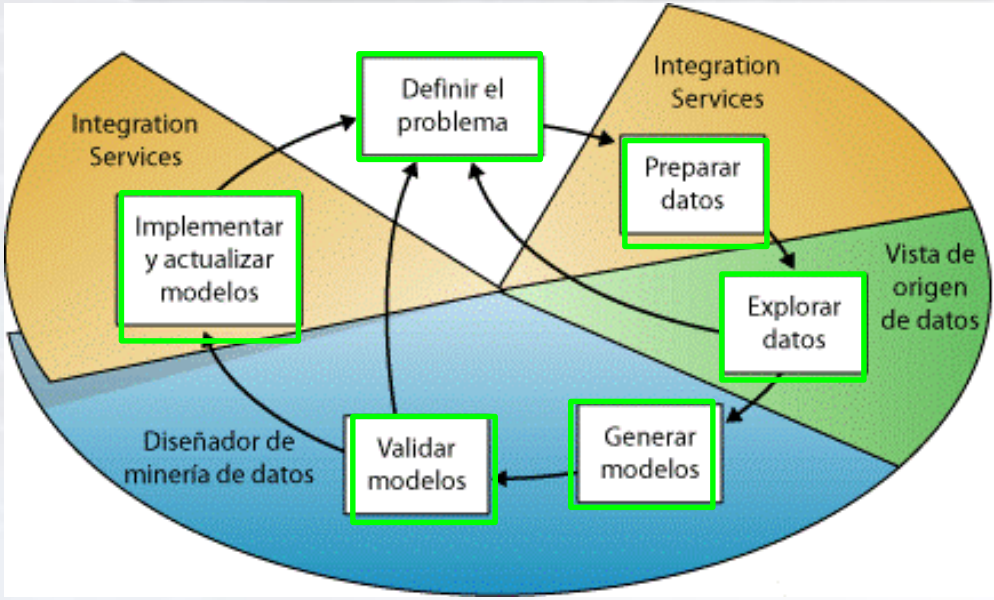
P5. Validar Modelos. -

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 172   1   0   1   1   0   1 |   a = 7
   1 172   0   1   1   2   0 |   b = 1
   1   0 138   0   0   1   0 |   c = 4
   0   1   0 151   0   0   1 |   d = 6
   0   0   0   0 173   2   1 |   e = 2
   2   0   2   1   1 153   0 |   f = 3
   1   1   1   2   1   0 213 |   g = 5
```

Matriz de confusión reportada al aplicar *PART*. Reporta 28 errores, por ejemplo, para el miércoles (fila f=3) 153 registros se clasificaron correctamente y 6 no.

```
Number of Rules  :      127


Time taken to build model: 0.13 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances        1172               97.6667 %
Incorrectly Classified Instances        28                2.3333 %
Kappa statistic                          0.9727
Mean absolute error                      0.0095
Root mean squared error                  0.069
Relative absolute error                  3.8951 %
Root relative squared error             19.7361 %
Total Number of Instances             1200
```

METODOLOGÍA DE LA INVESTIGACIÓN



Procesos para la generación de un modelo de MD
Fuente: https://tinyurl.com/y9s6tys6

PART generó 127 reglas a partir de los 1200 registros con un error medio cuadrático de 0,069. Apenas 28 registros se clasificaron incorrectamente, tal cual se detalló en la matriz de confusión.

P6. Implementar y actualizar modelos. -

El perceptron multicapa, con su algoritmo *backpropagation* funciona bastante bien, con 3 capas ocultas de 6 neuronas cada una, pues el error medio absoluto indica la calidad de la medida del modelo al ser apenas del 0,02 después de que se revisó 500 veces (épocas) cada uno de los 1200 registros. La diagonal de la matriz de confusión representada en la Ilustración 24, que mide el acuerdo inter evaluador para las variables nominales que en este caso son los meses en que se cobró, clasificó correctamente 1196 registros, lo que guarda concordancia con el resultado reflejado en la medida estadística del *Coeficiente de Kappa* que alcanza el 0,9964 sobre 1.

Las reglas de asociación con el algoritmo PART, también reportan datos interesantes, pues el error medio absoluto del modelo es apenas del 0,0095 una vez que se revisó las 1200 instancias que generaron un total de 127 reglas, el modelo clasificó correctamente cerca del 98% de las instancias, lo que se respalda con la matriz de confusión representada en la Ilustración 26 que mide el acuerdo inter evaluador para las variables nominales analizadas y concuerda con la medida estadística del Coeficiente de Kappa que alcanza el 0,9727 sobre 1. Adicionalmente, el algoritmo PART resulta menos complejo de actualizar por parte del personal de TI de la empresa eléctrica, en comparación con el perceptron multicapa dado que la cantidad de configuraciones que se requiere es menor.

DISCUSIÓN

El pronóstico de la demanda de energía eléctrica es un procedimiento sistemático que permite definir cuantitativamente la demanda futura procurando la exactitud de la información (Ariza Ramírez, 2013, p. 24), sin obviar las incertidumbres; los resultados de las validaciones, los cruces de información...

Los enfoques basados en estructuras de minería de datos generan resultados favorables, por ejemplo, en la predicción de los días de mayor pago... pero los algoritmos presentados pueden tener limitaciones como el hecho de que los modelos de redes neuronales o de reglas de asociación, según (Li & Wen, 2014) pueden no funcionar adecuadamente fuera de sus datos de entrenamiento o sí se generaliza o no, mucho más allá del rango de entrenamiento.

Las reglas de asociación han sido incluidas por la IEEE International Conference on Data Mining, entre los diez primeros algoritmos de minería de datos más influyentes en la comunidad de investigación (Wu et al., 2008, p. 2), en tanto que las redes neuronales son en concreto una de las estructuras más usadas en la predicción de consumos eléctricos (Ahmad et al., 2014).

Respecto al presente estudio, y más en particular sobre la construcción del archivo ARFF extraído a partir de la base de datos MySQL de la empresa eléctrica, y que contó con 1200 registros que corresponden a consumos eléctricos del sector residencial en el año 2015, en la práctica se debe contar por lo menos con registros de 10 años para el pronóstico de demanda de energía eléctrica.

CONCLUSIONES

No hay un modelo de minería de datos o combinación de algoritmos de aprendizaje automático único para todos los conjuntos de datos, por lo tanto, es esencial considerar caso por caso los aspectos discutidos en este documento, incluidos los datos disponibles y las propiedades de estos algoritmos, en favor de mejoras entre las cuales resalta el análisis de la eficiencia energética. Aunque la verdadera importancia del pronóstico de la demanda se incrementa en la medida que el cumplimiento de los objetivos trazados dependa lo menos posible del azar, incluso es recomendable que en el caso de los perceptrones multicapas se realicen simulaciones paramétricas que determinen combinaciones más precisas en cuanto a número de capas y neuronas por capas, disipando la posible incertidumbre sobre los resultados de las decisiones tomadas a partir de los modelos.


Make way!
I'm a data scientist!

REFERENCIAS

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, *33*, 102-109. https://doi.org/http://dx.doi.org/10.1016/j.rser.2017.04.095

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192-1205. https://doi.org/http://dx.doi.org/10.1016/j.rser.2017.04.095

Ariza Ramírez, A. M. (2013). *Métodos utilizados para el pronóstico de demanda de energía eléctrica en sistemas de distribución*. Universidad Tecnológica de Pereira, Pereira - Colombia. Recuperado a partir de https://tinyurl.com/y7akrz7z

Gönen, T. (1986). *Electric power distribution system engineering*. New York, New York, USA: McGraw-Hill.

John Lu, Z. Q. (2010). The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 693-694. https://doi.org/http://dx.doi.org/10.1016/j.enbuild.2014.02.005

Li, X., & Wen, J. (2014). Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews*, *37*, 517-537. https://doi.org/https://doi.org/10.1016/j.rser.2014.05.056

Microsoft. (2018). Data Mining Concepts. Recuperado 1 de agosto de 2018, a partir de https://tinyurl.com/yay5hjqt

Rosenblatt, F. (1961). *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Buffalo, NY: Cornell Aeronautical Lab Inc. Recuperado a partir de https://tinyurl.com/yb8qk6zz

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.

Van Heddeghem, W., Lambert, S., Lannoo, B., Colle, D., Pickavet, M., & Demeester, P. (2014). Trends in worldwide ICT electricity consumption from 2007 to 2012. *Computer Communications*, *50*, 64-76. https://doi.org/https://doi.org/10.1016/j.comcom.2014.02.008

Wang, Z., & Srinivasan, R. S. (2015). A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models. En *Winter Simulation Conference (WSC), 2015* (pp. 3438-3448). IEEE. https://doi.org/10.1109/WSC.2015.7408504

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4.ª ed.). Burlington, MA: Morgan Kaufmann. Recuperado a partir de http://www.cs.waikato.ac.nz/~ml/weka/book.html

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37. https://doi.org/DOI 10.1007/s10115-007-0114-2

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and buildings*, *75*, 109-118. https://doi.org/https://doi.org/10.1016/j.enbuild.2014.02.005

Uleam
UNIVERSIDAD LAICA
ELOY ALFARO DE MANABÍ

ESPAMMFL
ESCUELA SUPERIOR POLITÉCNICA
AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ

FACCI
FACULTAD DE CIENCIAS INFORMÁTICAS

Carrera de
COMPUTACIÓN

CIDE
Centro de Investigación y Desarrollo Ecuador

ULEAM
EMPRESA PÚBLICA DE PRODUCCIÓN Y DESARROLLO ESTRATÉGICO

Mantahost
Hotel

II Congreso Internacional de Tecnologías de la Información y Computación CITIC 2018

CITIC

29, 30 y 31 octubre 2018

Manta – Manabí
Sede:ULEAM