Procedimiento de Explotación de Información para la Identificación de Campos anómalos en Base de Datos Alfanuméricas

Horacio Kuna[1,2], German Pautsch[1], Aalice Rambo[1], Martin Rey[1], J.ose Cortes[1], Silvina Rolón.
[1]. Departamento de Informática,
Facultad de Ciencias Exactas Químicas y Naturales Universidad Nacional de Misiones.
Misiones, Argentina
[2]. Fac. de Ingeniería – Universidad Nacional de Itapúa, Paraguay
hdkuna@gmail.com

***Resumen*—La auditoría de sistemas dentro de las organizaciones desempeña una actividad de control y prevención sobre los sistemas, los cuales manejan uno de los activos más importantes que tienen las organizaciones que es la información. Para realizar estas actividades los auditores cuentan con una serie de técnicas y herramientas que los asisten, las Técnicas de Auditoría Asistidas por Computadora (TAACs). El presente trabajo muestra un procedimiento que utiliza técnicas de minería de datos que logran detectar campos considerados anómalos en una base de datos.**

***Abstract*—The audit of systems within organizations may perform a control and prevention on systems that handle one of the most important assets in the organizations, this asset is the information. To perform these activities the auditors have a number of techniques and tools that assist them, the Computer assisted audit techniques (CAATs). This paper shows a method that uses data mining techniques that detect anomalous fields in the database.**

Index Terms—Campos Anómalos, Bases de Datos, Datamining.

I. INTRODUCCION

El manejo de la información en las organizaciones en formatos digitalizados y electrónicos es un elemento en constante crecimiento, que lleva a almacenar grandes volúmenes de datos provenientes de diversas fuentes. Puede suceder que durante su captura o procesamiento se generen situaciones que los afectan volviéndolos anómalos, estos datos con valores extremos pueden afectar la calidad de la información que se utiliza tanto para las operaciones transaccionales como para las orientadas a la toma de decisiones. Por este motivo, la auditoría de los sistemas y de los datos que manipulan los mismos es una práctica no solo recomendable sino además aceptada por todos los integrantes de la organización, ya que establece mecanismos que ayudan a lograr mayor confianza en la información que se obtiene y procesa.

Dentro de las técnicas y herramientas que utilizan los auditores de sistemas existen diferentes propuestas. La MD (minería de datos), conocida como el proceso de extracción inteligente de información no evidente, pero presente en las bases de datos, es una de ellas pero su aplicación es aún incipiente.

Algunas técnicas de MD se encuentran orientadas a detección de outliers [1]. Un outlier es aquel dato [2] que, por sus características diferenciadoras en comparación a los demás datos contenidos en la BD, es sospechoso de haber sido introducido por otros mecanismos.

Los datos anómalos pueden crear distorsión en los resultados obtenidos al realizar cualquier tipo de análisis sobre los mismos. Sin embargo son menos frecuentes los estudios sobre la calidad de los datos, considerando a los outliers como posibles datos inaceptables, teniendo en cuenta como criterios de calidad la detección de datos anómalos, sucios o con ruido.

En cuanto a realizar trabajos específicos de MD existen propuestas que definen una serie de actividades tendientes a ordenar el proceso, entre ellos la metodología SEMMA [3] (Sample, Explore, Modify, Model, Assess), también aparece CRISP-DM (Cross-Industry Standard Processfor Data Mining) [4] y P3TQ [5] (Product, Place, Price, Time, Quantity).

Sobre la detección de outliers existen trabajos que definen una taxonomía de las anomalías detectadas en la búsqueda de outliers [6] en diferentes contextos como detección de fraude tanto en tarjetas de crédito [7] [8] como en teléfonos celulares [9], entre otros.

El problema que aparece en todos los trabajos que desarrollan algoritmos para la detección de outliers en Bases de Datos (BD) es que los mismos detectan las tuplas o filas que son consideradas anómalas y no detectan específicamente cuál es el campo específico de esa fila que tiene valores atípicos. Cuando se analizan grandes BD que contienen una importante cantidad de atributos, esta limitación de solo poder detectar las filas con valores anómalos es un inconveniente para el auditor, ya que debe realizar un análisis manual de cada uno de los campos de la fila sospechosa de contar con elementos que no son considerados normales. Un solo algoritmo no soluciona este déficit, por eso este trabajo propone un procedimiento que combina varios algoritmos para poder detectar específicamente cuál es el campo anómalo en una BD alfanumérica.

El presente trabajo plantea utilizar técnicas de MD, entre ellas *LOF* (Local Outlier Factor) [10] que es un algoritmo basado en la densidad, creado específicamente para detectar outliers y da como resultado un valor de un objeto *p* que representa el grado en que *p* es un *outlier*, de esta manera se logra identificar valores inconsistentes pudiendo mejorar la calidad de los datos.

El presente artículo propone un procedimiento el cual determina las filas y los campos dentro de un conjunto de

datos alfanuméricos que son *outliers*. Cabe destacar dos aspectos, primero que se trabaja con datos alfanuméricos obteniendo muy buenos resultados y segundo, que se logra identificar exactamente la fila y la columna donde aparece el *outlier*.

En la sección 2, Materiales y métodos, se describen los algoritmos utilizados y el procedimiento diseñado. En la sección 3, Resultados y discusión, se presentan los resultados obtenidos en la experimentación.

En la sección 4, Conclusiones, se identifican los principales logros del presente estudio.

En la sección 5, Referencias, se puede observar el compendio bibliográfico utilizado de referencia

II. Materiales y Métodos

El procedimiento propuesto utiliza una serie de pasos para determinar específicamente los campos (fila, columna/ registro, campo) que representa probablemente un dato anómalo; el proceso se aplicó sobre campos alfanuméricos.

A. Algoritmos de inducción

El proceso en primera instancia utiliza el algoritmo C4.5 para crear un árbol de decisión para determinar los campos significativos. Para ello el algoritmo realiza particiones en forma recursiva utilizando la estrategia "primero en profundidad" (*depthfirst*)[11]. En las pruebas sucesivas que realiza, busca aquellas cuyo resultado brindan la mayor ganancia de información. En el caso de atributos discretos la prueba considera una cantidad de resultados teniendo en cuenta el número de valores posibles que puede tomar el atributo. La detección de los atributos significativos dentro de la BD, aplicando algoritmos de Inducción como el algoritmo C4.5, permiten reducir el espacio de búsqueda de datos anómalos a solo aquellos campos que son relevantes en la BD, es decir, aquellos atributos que representan dentro del grupo de datos a los que aportan más información para clasificar al atributo *target* u objetivo, y poder de esta manera optimizar la performance y funcionamiento del procedimiento que detecta los campos anómalos.

Con este conocimiento el experto en el dominio podrá seleccionar los atributos a ser analizados en busca de *outliers* y descartar otros sin relevancia. (p.e. Nombre, Apellido, etc)

B. Algoritmo basado en la densidad. LOF (Local Outlier Factor)

Dentro del presente procedimiento hay una etapa para determinar la calidad de las filas aplicando LOF (*Local Outlier Factor*) [12], el cual pertenece al conjunto de técnicas basadas en densidad para la detección de *outliers*. Esta técnica hace uso de la estimación de densidad de los objetos, para ello, los objetos localizados en regiones de baja densidad y, que son relativamente distantes de sus vecinos, se consideran anómalos.

$$LOF_{MinPts}(x) = \frac{\sum_{y \in N_{MinPts}(x)} \frac{lrd_{MinPts}(y)}{lrd_{MinPts}(x)}}{|N_{MinPts}(x)|}$$

Cálculo de LOF (1)

Dada una instancia *x*, su *lrd* se define como la inversa de la distancia de alcanzabilidad promedio basada en la vecindad más cercana *MinPts* de la instancia *x*. Cuando la densidad de los vecinos de una instancia *x* es alta o, cuando su densidad es baja, entonces su LOF será grande y puede ser considerado un *outlier* [10].

El algoritmo LOF determina un factor local de o*utlier*, este factor puede tomar valores entre 0 e ∞, este valor es incorporado a cada tupla para su posterior análisis. Este algoritmo incorporado al procedimiento logra identificar la tupla que posee el campo con datos inconsistentes o con ruido. En particular, LOF es utilizado en el procedimiento para analizar la entrada (E) tomando siempre como referencia los valores existentes en las salidas (S) que corresponden al atributo definido como objetivo o *target*. Los valores MinPts son ingresados a la herramienta a través de la configuración de parámetros. La configuración de estos parámetros debe realizarse con el asesoramiento de un experto en el dominio así como la elección del valor de LOF resultante. Este último determinará el umbral por el cual un dato será o no un *outlier*.

C. Teoría de la Información

La teoría de la información nace como un modelo matemático enunciado en 1948 por Claude Shannon[13]. El sistema de comunicación propuesto posee una fuente que determina los mensajes a ser transmitidos, un transmisor que codifica el mensaje convirtiéndolo en una señal que se propaga por medio de un canal de transmisión. La señal llega al decodificador que la convierte nuevamente en el mensaje para el destinatario. Este mensaje puede ser idéntico al generado en el emisor, o similar, en el caso que se encuentre sometido el canal de transmisión a una fuente de ruido durante la transmisión del mensaje. La información se mide mediante la entropía, que es un término de la termodinámica que mide el nivel de desorden de un sistema; en teoría de la información, se refiere a la cantidad de información promedio que contienen los símbolos usados. Es decir, cuanto menor probabilidad de aparición tiene un símbolo, mayor es la cantidad de información que el mismo aporta.

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Entropía (2)

Como puede observarse n la función 2 de entropía de un set discreto de probabilidadesp_1 p_n. Shannon 1943.

Esta teoría aplicada en proceso de minería de datos [14] indica la posibilidad de trabajar los datos desde binomios del tipo "mensaje de entrada" (E) y "mensaje de salida" (S) para detectar los *outliers* en cada atributo.

Si consideramos el ejemplo de arrojar un dado (E), sobre esta acción resultante (S) debería existir un número entre 1 y 6. Si esto no es así estaríamos en presencia de un *outlier*. De esta manera, cuanto menor sea la probabilidad del par (E) – (S) analizado, más tendencia presentará a corresponderse con una inconsistencia.

Teniendo en cuenta lo anteriormente mencionado se utiliza LOF para analizar la entrada (E) tomando siempre como

referencia los valores existentes en las salidas (S) que corresponden al atributo definido como objetivo o target.

D. Procedimiento propuesto.

El procedimiento propuesto consta de cinco pasos los cuales abordan los conceptos vistos anteriormente. La aplicación de algoritmos de inducción en una primer instancia de acuerdo al atributo *target* u objetivo definido previamente.
Una vez identificado el conjunto de atributos representativos del set de datos, teniendo en cuenta los conceptos de teoría de información, por cada uno de ellos se componen pares entradas-salidas (E+S) con cada atributo y el atributo target identificado. Para evitar inconvenientes con los valores nulos, estos son reemplazados por una etiqueta "nulo", y se aplica LOF como algoritmo basado en la densidad para determinar que aquellos elementos que se encuentran en regiones de baja densidad se consideran anómalos.

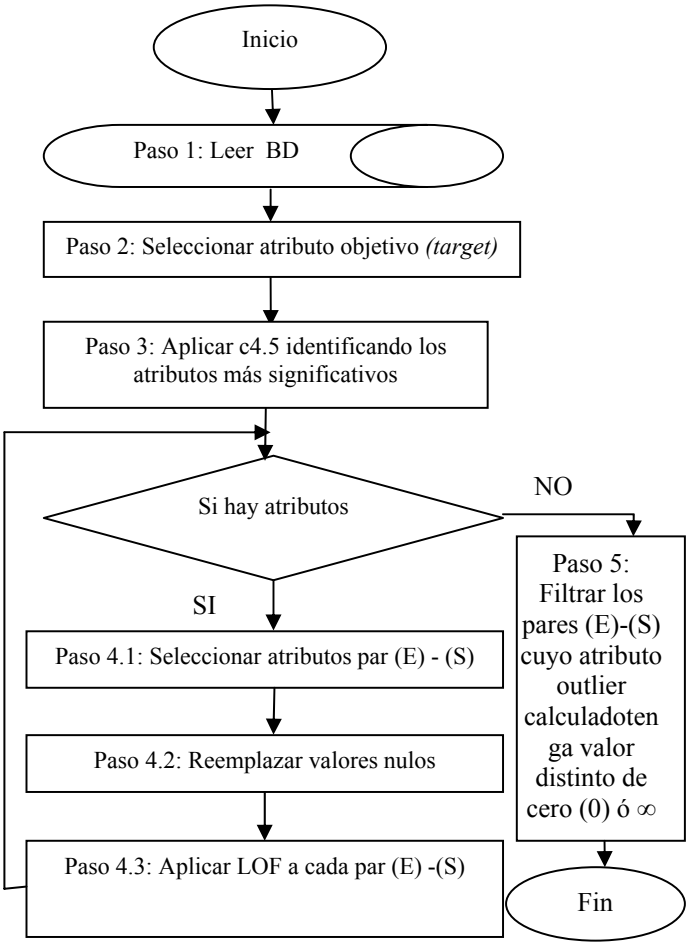

Fig. 1. Procedimiento de clasificación utilizando LOF para detectar outliers

A continuación se detallan cada una de las etapas que involucran al procedimiento descripto anteriormente, los mismos pueden verse en el diagrama de la figura 1.

- Paso 1: Se procede a tomar los datos del repositorio, para ello se utilizó la instancia Read CSV de la herramienta utilizada, Rapid Miner [17].
- Paso 2: Luego con la instancia Set Role se indica el atributo target u objetivo.
- Paso 3: Posteriormente con el operador Decision Tree se aplica el algoritmo c4.5 que deriva en el árbol de decisión que va a utilizarse como pre proceso para la selección de los atributos que son más relevantes para clasificar al target definido.
- Los pasos desde 4.1 a 4.3 se realizan por cada atributo seleccionado a partir del algoritmo de inducción.
    - Paso 4.1: Por cada atributo representado en el árbol se ejecuta el sub proceso de minería, el cual arma un *bin,* conjunto de datos (E)+(S) como se explicó anteriormente, tomando como entrada (E) el atributo seleccionado y como salida (S) el atributo *target* u objetivo. Esto se realiza en el flujo de minería con el operador *Select Attributes*.
    - Paso 4.2: Los datos que componen al atributo entrada son analizados en busca de valores nulos, los cuales de existir, en la instancia ReplaceMissingValues son reemplazados por una etiqueta para su posterior proceso.
    - Paso 4.3: Por último, se aplica LOF al bin (E)+(S) dando como resultado la generación de un atributo outlier
- Paso 5: Se filtran aquellos pares de datos bin (E)-(S) cuyo valor de outlier, al ser distinto de cero (0), nos indica la presencia de un dato anómalo que no aporta información y corresponde a ruido en referencia al target definido.

E. Experimentación

Los datos corresponden a las bases de datos de Hongos pertenecientes al repositorio "*Machine Learning Repository*" de la UCI (*University of California* - IRVINE) [15]. Es una BD nominal que posee una clasificación de hongos que permiten definir un atributo, objetivo o target que es el campo: *"clase"* sobre el cual se evaluan los demás atributos. La BD tiene 23 atributos y 8124 tuplas. En la tabla 1 se observan los atributos de la BD.

TABLA 1 ATRIBUTOS DE LA BD HONGOS

Forma _sombrero	**Superficie _sombrero**	**Superficie _sombrero**	**Magulladuras**
Olor	Tipo _membrana	Espaciado _membrana	Tamaño _membrana
Color membrana	Forma _tronco	Raiz_tronco	Sup_tronco _arriba_anillo
Sup_tronco_ debajo_anillo	ColorTronco arriba_anillo	ColorTronco _debajo_anillo	Tipo_velo
Color_ velo	Cantidad _anillos	Tipo_anillo	Color_esporas
Poblacion	Habitat	Clase (Target)	

Además se realizan diversas pruebas en bases de datos reales obteniendo similares resultados. El software para aplicar los algoritmos de minería es el *Rapid Miner* [16] y para los resultados se utilizó *Calc* de *Open Office* [17].
Cuando se experimenta con Bases de Datos reales que no responden a un tipo de distribución específico, como es este caso, la individualización previa de los campos considerados anómalos se dificulta, por este motivo con el objetivo de

validar los resultados del procedimiento propuesto se analizó en forma detallada la BD con un experto en hongos comestibles, quien detectó en forma manual los posibles outliers existentes en la BD, encontrándose un total de 59 campos que el experto consideró sospechosos de ser anómalos.

En el presente caso de estudio se ha configurado reemplazar en la entrada (E) los valores nulos por la etiqueta "nulo", de esta manera se mantiene la característica nominal de los datos.

Para la instancia *DetectOutier* (LOF) incorporada dentro del flujo de minería en la herramienta *Rapid Miner*, los valores de configuración para MinPts se definen como límite inferior en 10 y límite superior en 20 y el cálculo de la función de la distancia que se utiliza para el cálculo de la misma entre dos objetos es la distancia euclidiana. En la figura 2 puede verse el flujo de minería creado en el RapidMiner.

Al ejecutar el flujo de minería, LOF detecta los outliers presentes en el atributo seleccionado como (E) incorporando el atributo outlier calculado y colocándole el valor ∞ (infinito). Esto se puede observar en la vista de resultados activando la opción *Data View,* donde se puede verificar que un valor de *outlier* distinto de cero o infinito representará ruido en el atributo (E).

Como se mencionó anteriormente, cuanto menor sea la probabilidad del par (E) – (S) analizado, mayor es la posibilidad de que corresponda a una inconsistencia.

Se forman los "bin" [14] con los atributos de entrada (E) y el atributo target (S) y su cálculo de LOF representando la relación existente entre (E) y el elemento (S). En la teoría la información, cuando Shannon [13] hace referencia a la cantidad de información que aporta el elemento, la cual se corresponde con la probabilidad de que ese elemento aparezca en la salida donde E es el mensaje emitido y S es el mensaje recibido. En el presente proceso el elemento (E) con baja densidad con respecto al elemento (S) representa alta probabilidad de que represente ruido (elemento anómalo). Se utiliza LOF donde la densidad de ese elemento con respecto al atributo target es muy baja, por lo cual indica que es probable que se corresponda con ruido.

III. Resultados

El resultado de ejecutar el algoritmo C4.5 permitió definir 5 atributos significativos: *espaciado_membrana, forma_sombrero, forma_tronco, olor, tipo_membrana,* junto con el atributo target: *clase.*

En la BD de hongos al analizar el atributo *forma de sombrero*, se encontraron 4 registros con forma de sombrero cónica y clase venenoso identificados como outliers, cabe destacar que no hay otra combinación con forma de sombrero cónica.

También aparece el atributo "forma de sombrero" = "acampanado" con "clase" = "venenoso", los 4 únicos registros con esa combinación se registraron como *outliers*, todos los demás registros con forma de sombrero acampanada corresponden a la clase ingeribles. Esto se repite en el atributo "olor", donde el olor es un atributo con mucho peso (marcado por el árbol de decisiones generado en el procedimiento). Sin embargo, difícilmente un hongo que no tenga olor puede ser venenoso. Se detectan 10 casos entre los registros utilizados en esta prueba.

En el atributo "tipo membrana" aparecen 18 registros con la combinación "tipo membrana" = "adherida" y "clase" = "venenoso", el experto consideró esta combinación como anómala.

En la tabla 2, en la columna aciertos se determina en base a los outliers detectados y los existentes en la BD. Como se puede observar en el caso del atributo "forma tronco", los 12 campos detectados como *outliers* correspondían a datos nulos.

TABLA II. Outliers detectados en la Base de Datos de hongos

Atributo	Cant. datos	Outliers detecta-dos	Outliers existen-tes	Datos nulos	Aciertos
espaciado_ membrana	8124	8	8	0	100,00%
forma_sombrero	8124	12	8	0	100,00%
forma_tronco	8124	27	15	12	100,00%
Olor	8124	10	10	0	100,00%
tipo_membrana	8124	18	18	0	100,00%

Del total de 59 campos considerados *outliers* por el experto, el procedimiento detectó el 100% de estos campos.

IV. Conclusiones

El procedimiento especificado permite de manera sencilla identificar qué campos y qué atributos presentan algún tipo de inconsistencia.

El porcentaje de acierto es mucho mayor al esperado registrándose un 100% de detección de *outliers.*

El procedimiento precisa de un campo clasificador o *target* para realizar las combinaciones de (E) – (S).

En futuras líneas de investigación se plantea probar el comportamiento del presente procedimiento con algoritmos de clusterización.

V. Referencias

[1] Torr P.H.S. and Murray D. W.: Outlier Detection and Motion Segmentation. Sensor Fusion VI Volume: 2059, Pages: 432-44. Robotics Research Gorup, Department of Engineering Science, University of Oxford Parks Road, Oxford OX1 3PJ, UK. (1993)

[2] Hawkings, D.: Identification of Outliers.Chapman and Hall. London. (1980)

[3] SEMMA. 2008. http://www.sas.com/technologies/analytics/ datamining/miner/semma.html. Vigencia 15/09/08.

[4] CRISP-DM. 2008. http://www.crisp-dm.org/ Vigencia 15/09/08.

[5] Pyle, D.: Business Modeling and Business intelligence. Morgan Kaufmann Publishers (2003)

[6] Chandola V., Banerjee A., and Kumar V.: Anomaly Detection: A Survey. University of Minnesota. Pg 15-58. ACM Computing Surveys, Vol. 41, No. 3, Article 15. (2009).

[7] Bolton, R. And Hand, D.: Unsupervised profiling methods for fraud detection. In Proceedings of the Conference on Credit Scoring and Credit Control VII. (1999)

[8] Teng, H., Chen, K., and Lu, S.: Adaptive real-time anomaly detection using inductively generated sequential patterns. In Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. IEEE ComputerSocietyPress, 278–284. (1990).

[9] Fawcett, T. and Provost, F.: Activity monitoring: noticing interesting changes in behavior. In Proceedings of the 5th ACM SIGKDD International Press, 53–62. Conference on Knowledge Discovery and Data Mining. ACM (1999).

[10] Breunig M. Kriegel H., Ng R., Sander J. LOF: Identifying Density-Based Local Outliers. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.

[11] Quinlan J. R., C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California. EEUU.(1993).

[12] Hu T. and Sungs S. Y.: Detecting pattern-based outliers. Pattern Recognition Letters, vol. 24, no. 16, pp. 3059-3068. (2003).

[13] Shannon C. A Mathematical Theory of Communication. Reprinted with corrections from The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

[14] Ferreyra M. Powerhouse: Data Mining usando Teoría de la información. Octubre (2007).

[15] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.: últimavisita 20/05/2012.

A. *S*oftware Utilizado

[16] RapidMiner. Sistema Open Source para minería de datos. http://rapid-i.com/content/view/26/84/lang,en/ (18-02-2012)

[17] Open Office. Calc. Programa de Hoja de Cálculo. Open Source. http://www.libreoffice.org/features/calc/ (18-02-2012).

Horacio Kuna. Licenciado en Sistemas egresado de la Universidad de Morón, Master en Ingeniería del Software egresado del ITBA y la Universidad Politécnica de Madrid. Profesor Titular, Director del Departamento de Informática y del Programa de Investigación en Computación de la Fac. De Cs.Exactas Químicas y Nat. de la Universidad Nacional de Misiones-Argentina. Docente-Investigador de la Facultad de Ingeniería de la Universidad Nacional de Itapúa-Paraguay. Doctorando de Ingeniería en Sistemas y Computación, Universidad de Málaga-España, DEA aprobado, en desarrollo de tesis.

German Pautch. Licenciado en Sistemas de Información, egresado y actualmente docente de las cátedras Bases de Datos y Trabajo Final de la carrera Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones. Como no docente, responsable del Departamento de Sistemas de la Dirección de Tecnología para la Gestión de la Facultad de Ciencias Económicas, Universidad Nacional de Misiones

Alice Rambo. Ingeniera en Sistemas. Profesora de Informática. Docente a cargo de la cátedra Inteligencia Artificial y Sistemas Expertos y Modelos y Simulación de la Carrera de Licenciatura en Sistemas de Información. Docente Investigador del Departamento de Informática de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones

Rey Martín. Licenciado en Sistemas de Información. Docente-Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.

José Cortes. Alumno avanzado de la carrera de Licenciatura en Sistemas de Información. Investigador del Departamento de Informática. Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.

Silvina Rolon. Profesora de Inglés egresada del "Instituto Superior del Profesorado Antonio Ruiz de Montoya" y licenciada en Lengua Inglesa, egresada de la Universidad Nacional del Litoral. Actualmente, estoy trabajando en la Universidad de la Cuenca del Plata como profesora adjunta dictando las cátedras de Inglés I y II en las carreras Lic en Psicología, Lic en Nutrición, Lic en Psicopedagogía y Abogacía.