Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

MÉTODO DE PROCESAMIENTO DE LENGUAJE NATURAL Y TÉCNICAS DE MINERÍA DE DATOS APLICADAS A LA CLASIFICACIÓN DE INCIDENTES INFORMÁTICOS

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Diana Maribel Garcés Eslava

Código 20120535

Asesor

José Antonio Taquía Gutiérrez

Lima – Perú

Mayo de 2020

Recibido: 24 de septiembre del 2019
Aceptado: 19 de noviembre del 2019
doi: 10.26439/interfases2019.n012.4635

Método de procesamiento de lenguaje natural y técnicas de minería de datos aplicadas a la clasificación de incidentes informáticos

Diana Maribel Garcés-Eslava
20120535@aloe.ulima.edu.pe
Universidad de Lima. Lima, Perú

Resumen

El presente artículo plantea una metodología en la que se aplica el procesamiento de lenguaje natural y algoritmos de clasificación, haciendo uso de técnicas de minería de datos e incorporando procedimientos de validación y verificación de significancia, de acuerdo al análisis y selección de los datos, así como también de los resultados, con base en estadísticas de calidad de la información, lo que permite garantizar el porcentaje de efectividad en la construcción del conocimiento. Se utiliza como caso de estudio el análisis de incidentes informáticos en una institución educativa y una base de datos estandarizada de incidentes informáticos históricos recopilados de su área de mesa de servicio, la cual vincula con todos los procesos de tecnologías de la información y se centra en las necesidades de soporte para la realización de las actividades de los empleados. Mientras el requerimiento del usuario no sea resuelto en un tiempo adecuado, el impacto del incidente puede traer inconvenientes laborales de niveles variados, de tal manera que se hace difícil planificar o prevenir la resolución de los incidentes debido a la naturaleza imprevista de los mismos.

Palabras clave: *análisis de riesgos, incidentes informáticos, mesa de servicio, minería de datos, nivel de urgencia*

Abstract

Method of natural language processing and data mining techniques applied to the classification of computer incidents

This article presents a methodology that applies natural language processing and classification algorithms by using data mining techniques, and incorporating procedures for validation and verification of significance. This is conducted according to the analysis and selection of data and results based on quality statistical analysis, which guarantees the effectiveness percentage in knowledge construction. The analysis of computer incidents within an educational institution and a standardized database of historical computer incidents collected by the Service Desk area is used as case study. Such area is linked to all information technology processes and focuses on the support requirements for the performance of employee activities. As long as users' requirements are not fulfilled in a timely manner, the impact of incidents may give rise to work problems at different levels, making it difficult to plan or prevent incidents resolution due to their unforeseen nature.

Keywords: *risk analysis, computer incidents, service desk, data mining, emergency level*

1. Introducción

El nivel de avance tecnológico y el incremento exponencial en la capacidad de almacenamiento de los dispositivos informáticos, así como el volumen de información que manejan las organizaciones, hace que el proceso de análisis e interpretación de los datos sea complicado.

De los patrones de eventos y a través del uso de técnicas, algoritmos y mecanismos de validación de minería de datos, se puede obtener información sobre cómo surgen los incidentes informáticos y cómo tratarlos. Es importante que, en una organización, los equipos informáticos trabajen el mayor período posible, con el objetivo de garantizar la completa disponibilidad de los mismos y mantener la calidad de los servicios prestados.

La mesa de servicio, en una institución educativa, se encarga de planificar, organizar y suministrar la entrega de diversos servicios de tecnologías de la información, siendo el contacto principal entre el área de gestión de estos servicios y los usuarios. Además, se responsabiliza de atender todas las incidencias y peticiones de servicio, asegurándose de restaurar, lo más pronto posible, el servicio al cliente con bajo impacto, generando informes al respecto, en los cuales se identifica un catálogo de servicios donde se establecen los recursos y procesos que se necesitan para la ejecución de un servicio, permitiendo el ingreso de solicitudes, ya sea por cliente interno, usuario final o miembros del equipo del área de tecnologías de la información.

Actualmente, existen sistemas que se encargan de tratar los incidentes informáticos en la unidad de mesa de servicio, los cuales son herramientas relevantes para la documentación de estos problemas. Estas herramientas consideran ocurrencias desde la perspectiva de la gestión del servicio, teniendo en cuenta la configuración, los costos y el personal. Los fenómenos dinámicos de un entorno de mesa de servicio dan lugar a oportunidades para obtener conocimientos prácticos a partir de la información histórica, a través de la identificación de los posibles orígenes de las eventualidades y mediante la interpretación de los datos relacionados con la gestión de incidentes.

La minería de datos transforma datos en bruto a información útil, a fin de poder analizar los incidentes informáticos y monitorear fallas, haciendo uso de grandes bases de datos con la descripción detallada del problema resuelto; extrayendo así, rápidamente, el contexto del incidente. Para ello será necesario analizar y evaluar los datos históricos y descubrir correlaciones entre las variables que suelen intervenir con mayor frecuencia en la presentación de incidentes informáticos. El análisis y la predicción de la asociación son dos tareas importantes en la minería de datos, y van a representar dos objetivos principales: la exploración de datos para la comprensión y la construcción de modelos para la clasificación.

2. Herramientas para el procesamiento de datos

2.1 Weka e InfoStat

García *et al*. (2013) mencionan herramientas como Weka (*software* libre) e InfoStat (*software* comercial), las cuales fueron empleadas para el proceso de los datos y la construcción del modelo que realizaron. Sostienen que "Weka es utilizada para la transformación de datos, tareas de agrupamiento, regresión, clasificación y asociación e InfoStat se enfoca más al tratamiento de estadísticas, el estudio de datos multivariado y métodos para modelados estadísticos" (p. 108). Según los estudios que realizaron acerca del uso de minería de datos en el análisis de incidentes informáticos, los autores llegaron a las siguientes conclusiones: a) la minería de datos permite encontrar tendencias, patrones y comportamiento en una amplia base de datos que no suelen identificarse a simple vista; b) la técnica de asociación logra identificar asociaciones en un conjunto de datos, proporcionando reglas de asociación que aportan conocimiento para encontrar la relación entre las variables; c) el modelo otorgado por la técnica de minería de datos ayudará a la preparación de un plan de prevención para disminuir los tickets de incidentes informáticos generados constantemente en la organización.

2.2 ITIL (Biblioteca de Infraestructura de Tecnologías de Información)

En los últimos años, como menciona Baca y Vela (2015), se ha implementado un *software* web libre, ayudando en la mejora de la gestión de tecnologías de la información de incidencias, entendiendo la realidad del área y obteniendo información y estadísticas en tiempo real. Se busca, principalmente, utilizar estándares de calidad que permitan implantar un contexto de trabajo que involucre los procesos y servicios tal como propone ITIL. Al aplicar las buenas prácticas de ITIL, mejora el rendimiento en los servicios que ofrece la organización, así como satisfacer las necesidades del usuario y mejorar el rendimiento del personal para poder estandarizar los procesos y enfocarlos a la gestión de servicios de tecnologías de la información proporcionando calidad.

2.3 GLPI (Gestionnaire Libre de Parc Informatique)

Esta herramienta supone un plan de implementación que busca corregir los servicios para el proceso completo de las solicitudes que se presenten: recepción de incidentes, identificación, aplicación, asignación de la persona encargada de la resolución de los casos presentados, documentación y búsqueda de la mejor solución, ayudando así a restablecer el servicio y realizando un análisis exhaustivo, evitando casos o situaciones similares a las ya reportadas. Ballesteros, Hernández y Sánchez (2010) indican que, gracias a la obtención de información existente —como pueden ser datos estadísticos, resolución de incidentes históricos y reportes de auditoría externa e interna—, recopilada

mediante la búsqueda documental, las técnicas de localización de datos y el análisis de documentos, vamos a poder tener el conocimiento que favorezca al entendimiento de la funcionalidad en el proceso del control de incidentes informáticos.

3. Análisis de las técnicas de agrupamiento y asociación

Corso *et al.* (2014) aplicaron las técnicas de agrupamiento y asociación para identificar patrones y características en incidentes informáticos de equipos tecnológicos y realizaron el estudio en el marco de un laboratorio informático. La primera técnica mencionada propone un análisis exhaustivo de la información y los datos con el objetivo de identificar y dar solución a problemas de clasificación. La segunda técnica se usa para indagar e inspeccionar en un gran conjunto de datos, las reglas de asociación que evidencian las tendencias de las relaciones respecto a los datos de las entidades.

Fombona, Rodríguez y Barriada (2012) consideran que, dado los diversos orígenes sociales y culturales de los usuarios que integran los institutos educativos, tiende a haber una variedad de clientes y usos no uniformes en el manejo de los recursos informáticos. En el contexto de un centro educativo, en este caso la Universidad de Oviedo y en una de sus facultades (Formación del Profesorado y Educación), los autores estudiaron los incidentes informáticos frecuentes que recaen directamente en el personal administrativo y docente. Para este proceso eligieron una ficha de comunicación de incidentes con las siguientes variables, que después iban a recuperar en el sistema: fecha de inicio, docente que comunica la anomalía, lugar, descripción, contacto encargado de la incidencia, estado (resuelto, completado, en ejecución), control y seguimiento (pendiente de la siguiente actualización), comentarios u observaciones. Las anomalías que se presentaron fueron atendidas por el personal y alumnos con beca en informática que se hacían cargo de la resolución, en primera instancia, de las incidencias que eran detectadas en los equipos, así como también de la actualización de *software*, realización de inventarios y buen manejo de los recursos.

Después de utilizar este sistema de gestión propuesto, obtuvieron los siguientes resultados: a) el volumen de incidentes generados fue constante, pero hubo un incremento en el número de equipos informáticos; b) los usuarios supieron emplear mejor los dispositivos informáticos; c) la mayoría de incidentes se relacionaban con problemas de *software*, por lo que se propuso programar mantenimiento de los programas a utilizar y actualización de los servicios centrales; d) se incrementó el nivel de conocimiento de los usuarios respecto a la resolución de los problemas de forma automatizada dado que, a la par, aumentaron los recursos y aplicaciones informáticas; e) se implementó un protocolo claro y reducido, a fin de mantener una comunicación fluida con el coordinador y rapidez en la resolución de los incidentes presentados.

4. Comparación de modelos de clasificación

Barreno (2012) muestra un método para comparar el modelo de regresión logística y el de árbol de clasificación —ambos aportan en la clasificación para el estudio de deserción universitaria— y, de esta manera, establecer qué alumno podría ser clasificado como un desertor potencial, identificando previamente las variables más determinantes para el proceso correcto en la selección. Se utilizó el *software* Minitab 16 y Weka 3-7-2 para obtener los dos modelos. La autora consideró "desertor decisivo" al alumno que presente tres períodos consecutivos académicos sin registro de matrícula, no considerando en la muestra a los alumnos expulsados. Las variables explicativas con sus respectivos valores fueron las siguientes: colegio del que procede (1, colegio particular; 0, otro caso); efectividad examen de admisión (0;1); promedio ponderado acumulado (0;20); proporción de créditos aprobados (0;1); alumno entrega acta de compromiso (1, si el alumno presenta acta de compromiso; 0, en caso contrario); solicitud de recategorización (1, si el alumno ha solicitado recategorización; 0, en caso contrario); ingreso familiar (menos de S/ 2000; entre S/ 2000 y S/ 4000; más de S/ 4000). Se realizó el modelo de árbol de decisión, el cual permite el ingreso de variables de entrada tipo numérico y nominal para la categorización respectiva, es decir, acepta atributos cualitativos sin tener que pasar por una normalización o transformación de los datos. Para este modelo, se utilizaron los parámetros por defecto, excepto por la obtención de un árbol sin podar y las instancias mínimas que se aceptaron por hoja fueron 3.

Estas especificaciones fueron dadas para obtener un árbol más grande en el que se observó la discriminación en la mayor cantidad de variables. Este modelo obtuvo un 95,09 % de clasificaciones correctas utilizando gran parte de las variables: SR, PPA, AC, IF, EEA y PCA, pero esto no se recomienda, ya que indica un posible sobreajuste que afectaría al modelo cuando los datos sean distintos a los de entrenamiento, puesto que habría una generalización de los mismos. Por esto, se realizó nuevamente el árbol, pero utilizando solo 4 variables explicativas: SR, PPA, AC e IF, obteniendo un 94,62 % de clasificaciones correctas (Barreno, 2012). La aplicación de estos modelos de clasificación dio resultados parecidos, puesto que, con las dos técnicas, el análisis respecto a la deserción universitaria dio, al menos, un 94 %.

Barreno (2012) sostiene que, a diferencia del modelo de regresión logística, el árbol de decisión no se representa a través de una fórmula matemática, sino mediante reglas de decisión asociadas a las variables elegidas y la evaluación se realiza de acuerdo a la prioridad de cada regla: solo se realizan las evaluaciones que sean necesarias hasta obtener un clasificador para el alumno evaluado. Dicho autor concluyó que la forma de analizar la deserción estudiantil es diversa, no hay una forma para el análisis del mismo porque depende de los objetivos planteados para dicho análisis, teniendo en consideración el alcance y los recursos de tiempo y económicos a utilizar. Indica que la técnica del árbol de decisión permite obtener reglas claras, sin necesitar una transformación previa en los datos y permite realizar una discretización de datos cuantitativos. Asimismo, para la obtención de buenos resultados, se

deberá reformular el modelo realizando diversas pruebas, adecuando los parámetros, a fin de obtener un árbol de decisión que permita descubrir información relevante, evitando el subaprendizaje y sobreajuste de los datos.

5. Metodología

El sistema propuesto implicó el procesamiento de los datos de texto, en el cual un programa de código Python lee los datos de incidentes de un archivo de entrada, línea por línea, y escribe las palabras clave más frecuentes en el archivo de salida. El algoritmo procesa el incidente informático ingresado, tanto el título como la descripción del registro, y asociará las variables clave determinantes para luego encontrar una relación entre ellas y proporcionar una categorización de los incidentes por nivel de urgencia. Este procesamiento reduce el error de entrenamiento durante la clasificación de los datos. Esto, posteriormente, incide en diferentes aspectos como: ayudar en la identificación del posible origen del incidente, elaborar procedimientos y tareas de mantenimiento a realizar periódicamente, disminuir tiempos muertos, y aprovechar y uniformar el trabajo del área para la planificación de actividades. Los árboles de decisión fueron investigados para aprender de la experiencia pasada (fase de entrenamiento) y luego predecir sujetos para nuevas descripciones de incidentes (fase de prueba).

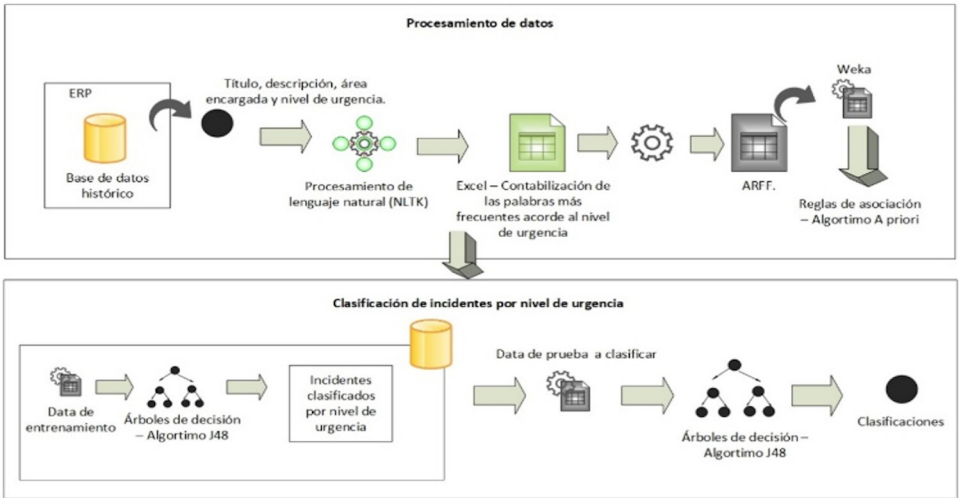

Figura 1. Diagrama de flujo del proceso para la clasificación de incidentes informáticos

Elaboración propia

La aplicación del procesamiento de lenguaje natural y las técnicas de minería de datos a implementar ayudan en la gestión de las bases de datos de incidentes informáticos para recuperar información relevante y, mediante el análisis estadístico y el aprendizaje de los

datos, otorgar una categorización a las incidencias reportadas continuamente, gracias a la identificación de relaciones entre los atributos y las características directamente asociadas al incidente, de tal manera que, al ser estudiadas y analizadas, aporten resultados favorables para el beneficio de la organización.

Se propone la técnica de minería de datos —reglas de asociación— para la búsqueda de relaciones y asociaciones entre los factores más influyentes de los incidentes informáticos. Adicionalmente, se utilizará el algoritmo J48 para la clasificación de los incidentes informáticos utilizando el método de árboles de decisión (ver figura 1). El proceso implica la normalización de los datos de texto, el título y la descripción del incidente, para extraer las palabras clave más recurrentes en una base de datos de incidentes informáticos. Esta normalización reduce el error de entrenamiento durante la clasificación de los datos. Un programa de código Python lee los datos de incidentes del archivo de entrada, línea por línea, y separa los términos clave más representativos de los incidentes informáticos, haciendo una contabilización de palabras clave por título y descripción del incidente.

De acuerdo a los incidentes informáticos reportados en la unidad de mesa de servicio y en base a la propuesta de investigación, se tendrán como objetivo los siguientes puntos: preparar los datos, identificar las variables y búsqueda; identificar asociaciones entre las variables; realizar la clasificación de incidentes informáticos.

Las variables del modelo fueron:

a) Palabras clave recurrentes: son las palabras clave con mayor recurrencia extraídas del título y descripción de todos los incidentes. i. Detalle pedido. ii. Detalle largo

b) Área encargada: el tipo de dato es nominal, tiene valores de "Desarrollo", "HelpDesk" y "Producción" de acuerdo al área al que será derivado el incidente informático.

c) Nivel urgencia: es el atributo más importante pues es la variable base para realizar la clasificación. Es de tipo nominal y tiene 3 valores: 100 (incidente de urgencia alta), 200 (incidente de urgencia media) y 300 (incidente de urgencia baja). Determinar la urgencia de los incidentes de tecnologías de la información sobre la herramienta de gestión es un elemento importante en la atención y gestión de incidentes, puesto que su clasificación permitirá dar relevancia a los incidentes que por su prioridad requieran una solución más pronta.

En base al conjunto de incidentes informáticos pasados, se clasifican los nuevos incidentes y se les asigna un tipo de urgencia de acuerdo a los factores tomados en consideración. Estos niveles de urgencia, en la clasificación de los incidentes informáticos, se toman en cuenta a fin de investigar cuál de ellos tiene un mayor impacto desde la visión de la organización, para que la unidad de mesa de servicio determine el orden deseado para la resolución

de los mismos y se enfoque en lo que es esencial y fundamental para evitar la discontinuidad del negocio.

Para determinar la urgencia, ITIL propone dos conceptos: a) impacto (métrica del efecto que causa en el proceso del negocio); b) criticidad (métrica del tiempo que tiene que pasar para que el impacto en el negocio sea considerado significativo). El nivel de urgencia debe definirse con un valor; también se puede establecer y dar un significado a cada nivel. En una organización los incidentes informáticos pueden ser clasificados en tres niveles: alto, medio y bajo.

6. Procesamiento de lenguaje natural (NLTK)

En primer lugar, se aplicó el método automático NLTK para el procesamiento de datos, a fin de encontrar palabras y expresiones características de un texto, es decir, palabras clave recurrentes en una base de datos de 2000 incidentes informáticos. Mediante la distribución por frecuencias se identificaron automáticamente las palabras más frecuentes de un texto; se utilizó el método "FreqDist" para encontrar las 50 palabras más recurrentes en el título y en la descripción de los incidentes. Se obtuvieron 150 palabras, de las cuales se escogieron las 43 más representativas que ayudaron a leer el título y la descripción del incidente. Dichas palabras clave fueron contabilizadas y clasificadas de acuerdo con el nivel de urgencia correspondiente.

Los campos necesarios para la extracción de datos se obtuvieron utilizando código Python para contar las palabras clave de la información extraída en la base de datos. El programa lee los datos de incidentes en un archivo de entrada en formato .arff, línea por línea, y escribe la clasificación del nivel de urgencia respectivo en el archivo de salida. Para ello, recoge el título y la descripción del incidente de cada línea, pasando cada uno de ellos a una función de conteo, enviando así esa contabilización, la cual devuelve como salida el nivel de urgencia alto, medio o bajo, según corresponda, y, a continuación, poder realizar el entrenamiento.

La función de conteo detecta palabras clave y contabiliza cuántas de cada una ocurren en el título y en la descripción. Estos conteos son enteros pequeños —la mayoría de ellos son cero—, puesto que los atributos tienen una o más palabras clave asociadas.

Una vez normalizados los datos, estos 2000 registros ingresados, incidentes históricos del año 2015 al 2018, fueron almacenados en un archivo con extensión .arff y presentaron los siguientes atributos:

- Palabras recurrentes: 43 palabras clave.
- Área encargada: 553 incidentes se derivan al área de Desarrollo, 1438 al área de HelpDesk y 9 al área de Producción.

– Nivel de urgencia (*class*): 527 incidentes de urgencia alta, 702 de urgencia media y 771 de urgencia baja.

A partir de la selección de cada uno de los atributos, se indica la cantidad de registros que hay en la muestra, de acuerdo a cada valor del atributo. Dado que el conjunto de datos que se está utilizando tiene el mismo conjunto de atributos en cada fila, es adecuado para el formato .arff.

7. Reglas de asociación – Algoritmo *a priori*

La técnica se enfoca en la generación de información en forma de reglas de asociación que tienen lugar comúnmente dado un conjunto transaccional determinado. Se basa en encontrar información útil desconocida en una base de datos y tan pronto como se generen conjuntos frecuentes, permite encontrar reglas de acuerdo a la confianza mínima propuesta por el usuario, dándoles prioridad a las que presentan una confianza mayor o igual a la indicada (Sharma y Bathia, 2016).

La minería de asociación busca relaciones interesantes en un gran conjunto de elementos de datos.

Supongamos que sea I = {I1, I2 ..... Im}, un conjunto de *m* atributos distintos, *T* la transacción que contiene un grupo de elementos, y *D* una base de datos con registros de transacciones únicos.

Una regla de asociación es un esquema en forma de X => Y, donde X, Y son conjuntos de elementos llamados *itemsets*.

Sean "x", "y" dos conjuntos de elementos diferentes en la transacción T, entonces la regla de soporte(s) de asociación se define como la relación de registros que contienen "x", "y" con el total de registros. Los dos umbrales se denominan "soporte mínimo" y "confianza mínima", respectivamente.

Parámetros iniciales: base de datos transaccional con 2000 registros, parámetro de apoyo mínimo con valor 1 y confianza mínima: 90 %

Pasos realizados: a) identificar conjuntos de elementos frecuentes usando el método a priori; b) parametrizar los índices de apoyo mínimo y confianza mínima; c) generar las reglas de asociación en base a los límites indicados para cada índice; d) obtener una lista de elementos con 3 variables por cada uno de los registros. Para este caso, se seleccionó el *lift* como el criterio para determinar la validez de las reglas, con un valor mínimo de 1, que se calcula como la confianza de la regla dividida por el soporte del consecuente de la regla. Se encontraron 7 recomendaciones y relaciones dentro de los registros ingresados. El *conf* es el porcentaje de acierto o confianza

que hace relación a las variables elegidas. El algoritmo se inició con el límite superior del soporte, fue disminuyendo en forma incremental y se detuvo al llegar al número especificado de reglas o cuando alcanzó el límite inferior del soporte (ver tabla 1).

Tabla 1

Aplicación de las reglas de decisión

Regla	Descripción	Evaluación
Title=transacción_pendiente 52 ==> category_tree_ value=desarrollo 52 conf: (1) Alto (100)	Si el contexto del incidente informático está relacionado a una transacción y, más aún, si se encuentra en estado "pendiente", debe derivarse al área de Desarrollo para que analicen dicha transacción y vean la causa del por qué el pago no se concretó. Es un escenario completo, en el cual el incidente afecta a activos considerados de impacto mayor que influyen directamente a los objetivos de la institución educativa. Estos incidentes deben tener respuesta inmediata.	Coherente: debido a que el valor de la elevación es 1, quiere decir que la ocurrencia no es aleatoria, sino que se debe a una cierta relación entre ellos.
Title=transacción_pendiente 52 ==> category_tree_ value=desarrollo 52 conf: (1)	Si el contexto del incidente informático está relacionado a una transacción y, más aún, si se encuentra en estado "pendiente", se determina que es un incidente de urgencia alta. Es necesario tener en cuenta el área a la cual será derivada el incidente para que la regla sea coherente.	Insuficiente: debido a que el valor de la elevación es 0,95, quiere decir que la ocurrencia es aleatoria.
Class=cien 71 ==> category_tree_ value=desarrollo 67 conf: (0,94)	No es una regla muy coherente debido a que no indica el título del incidente informático a tratar, por lo que no se puede determinar la categoría y la urgencia del mismo.	No coherente: debido a que el valor de la elevación es 0,94, quiere decir que la ocurrencia es aleatoria.

Elaboración propia

8. Árbol de decisión – Algoritmo J48

El algoritmo para construir el árbol de decisión se logra a partir de los siguientes pasos:

a) Identificar y comprobar los casos base.

b) Para cada atributo *a* (palabras clave determinadas previamente), encontrar la ganancia de información normalizada por cada división de *a*.

c) Dejar que *a_best* sea el atributo con la ganancia de información normalizada más alta, en este caso, el atributo clase de acuerdo al nivel de urgencia alto, medio y bajo.

d) Crear nodos de decisión que dividan a *a_best*.

e) Repetir estos nodos en las sublistas obtenidas por la división de *a_best* y agegar los nodos como hijos de *nodo*.

f) Entre los atributos explicativos se encuentran las variables cualitativas {x1, x2…xn} (palabras clave). Si una nueva instancia de evaluación es tal que su atributo x2 es menor o igual a c2 (números binarios de acuerdo a la contabilización de palabras), entonces clasificar la instancia como perteneciente a la categoría A de la clase de interés.

El algoritmo J48 se basa en un método para la ejecución de las funcionalidades propias de la técnica de árboles de decisión, en el cual lee el archivo en formato .arff ingresado con la data total, después procesa los atributos significativos (identificando palabas clave desde el texto del título y de la descripción de los incidentes, realizando la clasificación y contabilización de dichas palabras clave asignándolas a una categoría según su nivel de urgencia), construye el modelo, muestra el árbol generado para clasificación y, por último, indica el porcentaje de instancias correctamente clasificadas e instancias incorrectamente clasificadas. El J48 es supervisado y aprende de cada registro al ir generando el árbol de clasificación.

Se logró obtener una probabilidad de 80,2 % de instancias correctamente clasificadas, lo que indica que el algoritmo de predicción tiene un porcentaje adecuado de confiabilidad que garantiza que la técnica de árboles de decisión evite el sobreajuste y generalice las reglas de decisión identificadas, evitando un aumento de rendimiento, tanto en el conjunto de datos para la prueba, como para los de entrenamiento.

El árbol realiza la clasificación de las instancias de acuerdo a una clase con 43 categorías (palabras clave); los atributos que permiten realizar la clasificación son de tipo cuantitativo y la regla de decisión asociada es una comparación con respecto a un valor constante, ya que se compara el valor del atributo con cada una de las categorías que la conforman. El atributo de clasificación class (nivel de urgencia) es el atributo respuesta. Fueron clasificadas correctamente 80,2 % de las instancias.

A continuación, algunas clasificaciones obtenidas al emplear el árbol de decisión:

a) Si la palabra clave "matrícula" se encuentra en el título o descripción al menos una vez y la palabra clave "regularización" también figura, el nivel de urgencia va a ser bajo; pero si solo figura "matrícula", va a clasificarse como un incidente con nivel de urgencia medio. Esto quiere decir que la incidencia ingresada para la regularización de una matrícula en el sistema no tiene un impacto mayor en la organización, es decir, no es prioridad.

b) Si la palabra clave "cobro" figura en el título o descripción al menos una vez y la palabra "adicional" también se presenta, el nivel de urgencia es alto; pero si la palabra "adicional" no figura y la palabra clave "curso" se encuentra al menos una sola vez, entonces estaría clasificado en nivel de urgencia bajo. Esto quiere decir que la incidencia ingresada, debido a que se generó un cobro adicional en la tarjeta del alumno al realizar una matrícula, tiene un impacto mayor, puesto que se tendría que confirmar el doble cobro en la tarjeta del alumno y hacer la devolución, si se requiere.

Observando los resultados que se generaron por la técnica de reglas de asociación, se apunta a que estas puedan, del mismo modo, identificarse en el árbol de decisión, tal como se dio respecto a las reglas coherentes mencionadas anteriormente. Entonces, se puede decir que las reglas fueron importantes y efectivas de acuerdo a las clasificadas por nivel de urgencia por el algoritmo J48. Se obtuvo un porcentaje de 80,2 %, considerado como un porcentaje válido, por que se están generando reglas de ganancia de información y, a veces, el usuario puede utilizar otras palabras para expresar un problema. Para mejorar la técnica sería necesaria una estandarización de los datos, una manera de hacerlos más claros; para ello se debería seguir usando un procesamiento de lenguaje natural que analice y desmenuce las palabras, permitiendo hallar resultados importantes para la clasificación.

9. Validación

Para la validación y verificación de los datos históricos, se aplicarán técnicas de validación usando distintos validadores, con la finalidad de contrastar la precisión de los datos de entrenamiento con la data de prueba. Se pretende determinar si el comportamiento de la clasificación respecto a los incidentes informáticos presenta una alta confiabilidad en los datos propuestos.

Tabla 2

Matriz de confusión del árbol de clasificación

	Valor predicho		
Valor real	Alto	Medio	Bajo
Alto	390	30	107
Medio	45	508	149
Bajo	24	41	706

Elaboración propia

La matriz de confusión, generada a partir de la técnica de validación cruzada, *folds:10*, muestra el número de aciertos representado en los valores de la diagonal; los que están alrededor, serán los errores. Este método ayudará a comparar los resultados y a validar que estos no sean obtenidos al azar, pues los resultados de la clasificación de los datos deberán ser similares a los obtenidos en el entrenamiento con el algoritmo J48. El tomar valores aleatorios y analizar los resultados, permitió posicionarse en un escenario de prueba en el cual se pudieran validar los datos obtenidos por la predicción y clasificación del modelo, lo que facilitó obtener un porcentaje promedio por cada prueba del subconjunto de datos.

En total fueron 10 pruebas que tuvieron una eficacia del 74,6 % como resultado promedio final. Debido a que se tiene gran cantidad de palabras clave y los valores tienden a ser generalizados, el porcentaje de confiabilidad de las instancias correctamente clasificadas disminuyó con las técnicas de validación cruzada en comparación al árbol podado obtenido anteriormente.

Para cada uno de los valores que toma el atributo de clase, en este caso, el nivel de urgencia, el porcentaje de instancias con ese valor son correctamente predichas (TP: verdaderas positivas) y el porcentaje de instancias con diferentes valores son incorrectamente predichas a ese valor, aunque tenían otro (FP: faltas positivas). Las columnas de precisión y recobro son los validadores relacionados con los dos anteriores, que ayudan a probar que los resultados de precisión (fracción de instancias relevantes entre las instancias recuperadas) y recobro (fracción de instancias relevantes que se han recuperado sobre la total cantidad de instancias relevantes) son iguales en el análisis de los datos ya normalizados.

Tabla 3

Resultados de la validación cruzada

Clase	Precisión	Recobro	Medida F
100	0,796	0,653	0,717
200	0,702	0,691	0,738
300	0,694	0,860	0,768
Peso promedio	0,755	0,746	0,744

Elaboración propia

En la matriz de confusión generada a partir de la técnica de validación cruzada, *folds:10*, se obtiene el número de aciertos representado en los valores de la diagonal y el resto vendrían a ser las instancias erróneas. De los 527 incidentes informáticos con urgencia alta, 344 fueron bien clasificados y 183 no. Por otra parte, de los 702 incidentes informáticos con urgencia media, 485 fueron correctamente clasificados y 217 no. Asimismo, de los 771

incidentes informáticos con urgencia baja, 108 no fueron bien clasificados. Podemos decir que respecto a la serie de resultados que nos mostró la técnica de clasificación sobre los datos, contrastándolos con los de la técnica de validación cruzada, ayudan a comparar los resultados y validar que estos no son obtenidos al azar, pues los resultados de los datos son competitivos, uno con otro, teniendo como diferencia solo un 5,6 %.

Para la validación y verificación de los datos, se utilizó un algoritmo realizado en Python en el cual procesa, tanto la técnica de validación cruzada, como las métricas de exactitud y precisión, partiendo los datos en 10 iteraciones para el nivel de urgencia alto (100), medio (200) y bajo (300) por separado, logrando así determinar la confiabilidad en los datos. De igual manera, para los validadores de precisión y exactitud, se aplicaron dos indicadores propios de la técnica de árbol de clasificación: la medición Gini, que determina la probabilidad de una muestra aleatoria que clasifica correctamente si seleccionamos aleatoriamente una etiqueta según la distribución en uno de los nodos (ramas), y la entropía, que calcula la ganancia de información haciendo una división y mide la reducción de la incertidumbre sobre la etiqueta (nodo).

Se entiende que ambas son diferentes técnicas para dividir el nodo en modelos basados en el árbol de decisión. La mayoría de veces el rendimiento de un modelo no tiende a variar mucho si se usa Gini o entropía, pero en términos de computación, se puede decir que la medición de entropía toma más tiempo ya que incluye la función *Log*.

Se dividió aleatoriamente el conjunto disponible de registros en dos partes, uno para entrenar; otro para validar. El modelo de validación se ajusta al conjunto de entrenamiento, dando como resultado un modelo ajustado usado para predecir las respuestas para las observaciones en dicho conjunto.

El conjunto de datos para entrenar se usó a fin de adaptar, tanto el conjunto de validación, como el modelo para estimar el error de la predicción, para la selección correcta del modelo; el conjunto de prueba es usado de igual forma para evaluar el error de generación del modelo final elegido. Entonces, el error del conjunto de pruebas del modelo final subestimará el verdadero error de la prueba. Se dividió el 80 % de los datos para entrenamiento y el 20 % para la validación y prueba. El error de la prueba se logra estimar de acuerdo al promedio que proporcionan los validadores. Esta sección envuelve una división en un conjunto de observaciones dentro de 10 grupos de igual tamaño, siendo el valor de la media de error cuadrática de 0,3465 MSE; se computan las observaciones en *held-out* partes.

Tabla 4

Resultados por nivel de urgencia

	Nivel de urgencia					
	Alto (100)		Medio (200)		Bajo (300)	
K folds	Gini	Entropía	Gini	Entropía	Gini	Entropía
1	1	0,875	0,975	0,878	1	0,927
2	0,950	0,875	1	0,878	1	0,927
3	0,975	0,875	1	0,878	1	0,950
4	0,900	0,875	0,950	0,900	1	0,950
5	0,975	0,875	0,975	0,900	0,975	0,950
6	0,950	0,875	1	0,900	1	0,950
7	0,975	0,875	0,975	0,900	1	0,950
8	0,100	0,875	0,950	0,878	1	0,950
9	0,950	0,875	1	0,878	1	0,949
10	0,900	0,875	0,975	0,878	1	0,949

Elaboración propia

10. Resultados

Como se observa en los resultados se obtuvo el coeficiente de Gini y de entropía, sumando las diferencias por pares entre todos los valores y normalizando los atributos numéricos, haciendo que el valor mínimo y máximo del atributo sean cero y uno, respectivamente, para una transformación de escala. Se observa que todos los valores son iguales en el conjunto, puesto que la probabilidad está entre 0,9 y 1, lo cual indica que la probabilidad de que existan instancias correctamente clasificadas entre el total de registros es alta para el nivel de urgencia alto, de acuerdo a los incidentes informáticos que se puedan presentar en un futuro. Asimismo, se obtuvieron los valores para el índice de entropía con un promedio de 0,875 mostrando que la mayoría de los datos han sido agrupados por muestras en las clases a las que pertenecen, ya que maximiza la pureza de los grupos, tanto como es posible, cada vez que se crea un nuevo nodo del árbol de clasificación.

Mediante el código Python se obtuvieron los porcentajes de precisión y exactitud para los incidentes clasificados con nivel de urgencia alto, obteniendo un buen desempeño, con un 85 %, en el cual se puede observar que los números están relacionados linealmente; para los incidentes clasificados con nivel de urgencia medio y bajo, se alcanzó un valor de 65 %, aproximadamente. De igual forma, se utilizó la función *accuracy score* y se comparó con el índice de Gini que obtuvo entre 0,975 a 1, determinando la exactitud que se refiere a la proporción de las instancias correctamente clasificadas, para el nivel de urgencia medio y bajo, entre el total de los registros; del mismo modo con el índice de entropía que calculó el aporte de las variables utilizadas dando a conocer que, al no usar gran cantidad de variables (palabras clave), permite obtener un buen desempeño.

Estas dos métricas se utilizaron para el conjunto de prueba y esta función calculó la precisión, analizando la fracción (predeterminada) o el recuento (*normalize=False*) de las predicciones correctas. Dado que *normalize==True* devuelve las muestras clasificadas correctamente (*float*), de lo contrario devuelve el número de muestras incorrectamente clasificadas (int).

El mejor rendimiento fue 1 con *normalize==True* y el número de muestras con *normalize==False*. Esto indica que, el conjunto de etiquetas predicho para una muestra, coincide exactamente con el conjunto correspondiente de etiquetas en la matriz de indicador de cada etiqueta/matriz dispersa de datos, y que todo el conjunto de etiquetas pronosticadas para una muestra coincide estrictamente con el conjunto verdadero de etiquetas tal que la precisión es más próxima a 1.0.

Dado que se tuvo como objetivo utilizar un sistema para la correcta clasificación de los incidentes informáticos, de acuerdo al nivel de urgencia alto, medio o bajo, se realizaron pruebas en las cuales los resultados experimentales, al utilizar la técnica de *cross validation* (validación cruzada) con 10 iteraciones y obteniendo los índices de Gini y de entropía, mostraron que todos los árboles de decisión alcanzaron un porcentaje aceptable de exactitud y precisión, teniendo, además, beneficios de velocidad y buen rendimiento.

Uno de los problemas para la minería de datos, en un enfoque basado en palabras clave, es que se contaba con una gran cantidad de atributos que la mayoría de métodos de minería no podían manejar, por lo que se pudo analizar y probar el rendimiento disminuyendo las variables de 50 a 43 atributos, logrando mayor precisión predictiva y menor error cuadrático. Asimismo, los resultados indican que el J48 produce una solución rápida utilizando la técnica de *pruning*, puesto que la construcción del modelo tardó 0,1 segundos, obteniendo un 80,2 % de instancias correctamente clasificadas.

11. Conclusiones

Para efectos del trabajo, se propuso un análisis de los incidentes de tecnologías de la información utilizando datos de una institución educativa. En la actualidad, no se identifica correctamente la prioridad de los incidentes entrantes y se gasta mucho tiempo en resolver el problema y restaurar los servicios, dada la clasificación manual de los mismos.

Dicho esto, se propuso clasificar el título y la descripción de un incidente usando recuentos de palabras clave y algoritmos de clasificación, tales como reglas de asociación y árboles de decisión, con el objetivo de determinar la mejor categorización de los incidentes con un buen porcentaje de confiabilidad, velocidad y precisión. El modelo entrenado analizó el título y la descripción para predecir las posibles variables que podrían estar incluidas en los incidentes. Todo esto implicó la preparación de datos, selección de métodos y exploración de datos.

Los árboles de decisión alcanzaron un 80,2 % de precisión, por lo tanto, es una precisión aceptable, teniendo en cuenta los beneficios de velocidad y mejora en la automatización del proceso. Se mantuvo el enfoque basado en palabras clave para la clasificación del árbol de decisión de los incidentes informáticos de una institución educativa.

Uno de los problemas de este enfoque fue el número elevado de atributos, dado que los métodos de minería no los podían manejar. Sin embargo, se logró analizar y resumir los datos a 43 atributos con más recurrencia en la base de datos. Es así que el modelo propuesto es útil como un sistema de clasificación para predecir el tema de incidentes de tecnologías de la información en aplicaciones del mundo real.

12. Recomendaciones

En el futuro, sería de interés considerar otros algoritmos de aprendizaje automático, tal como la técnica de regresión lineal o análisis de conglomerados, y comparar el resultado en cuanto a precisión y exactitud con la técnica de árboles de decisión. Además, se necesitaría un sistema que incorpore todo el procesamiento, incluido un modelo desarrollado que enlace directamente las incidencias informáticas entrantes, para que una vez que se ingrese el problema, este sea clasificado automáticamente según el nivel de urgencia correspondiente, así pueda ser atendido de acuerdo a la prioridad determinada de manera automática.

El sistema también se puede vincular a una base de datos de conocimiento que ayudaría a resolver incidentes futuros basados en como se resolvieron problemas similares. El modelo de conocimiento generado brinda la clasificación adecuada de los incidentes de acuerdo a su nivel de urgencia alto, medio y bajo, lo que permitiría la elaboración de un plan preventivo que disminuya la generación de solicitudes y el tiempo de resolución del mismo, para actuar eficientemente frente a las nuevas incidencias que se presentan. Asimismo, esto ayudaría a que, posteriormente, se incentiven diferentes aspectos en una mesa de servicio, como la ayuda en la identificación del posible origen del incidente, la elaboración de procedimientos y tareas de mantenimiento a realizar periódicamente, y aprovechar y uniformar el trabajo del área para la planificación de actividades.

Como trabajo futuro se planea incluir nuevas variables que permitan una clasificación más personalizada de acuerdo al incidente, tomando en cuenta otros criterios, tal como el impacto económico que genera el problema a la organización; el período del año en el que se presenta mayor cantidad de incidentes y en qué situaciones de la institución educativa, por ejemplo, en época de matrícula, en las cuales se deba clasificar con mayor nivel de prioridad determinados incidentes para el buen rendimiento de los servicios; el rol de la persona que ingresa el requerimiento, dado que existe una estructura del personal de la organización que debe ser considerada para fidelizar al usuario brindándole servicios de calidad y priorizando sus solicitudes, en caso sea necesario; etc.

De igual forma, sería conveniente identificar la causa raíz de acuerdo a patrones concurrentes de la clasificación propuesta respecto a las palabras clave obtenidas de las incidencias, disminuyendo los tiempos en la resolución e identificando características relevantes en la presentación de los mismos.

Referencias

Baca, Y., y Vela, G. (2015). *Diseño e implementación de procesos basados en ITIL v3 para la gestión de servicios de TI del área de Service Desk de la Facultad de Ingeniería y Arquitectura* (tesis de licenciatura). Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de Lima.

Ballesteros, J., Hernández C., y Sánchez, S. (2010). *Propuesta para la mejora del proceso de control de incidencias dentro de una mesa de ayuda bajo el ciclo Deming* (tesis de licenciatura). Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas, Instituto Politécnico Nacional.

Barreno, E. (2012). Análisis comparativo de modelos de clasificación en el estudio de la deserción universitaria. *Interfases*, (*5*), 45-82. doi:10.26439/interfases2012.n005.149

Corso, C., García, A., Ciceri, L., y Romero, F. (2014). Minería de datos aplicada a la detección de factores para la prevención de incidentes informáticos. XVI Workshop de Investigadores en Ciencias de la Computación (pp. 168-172). Universidad Tecnológica Nacional. Recuperado de http://sedici.unlp.edu.ar/bitstream/handle/10915/41982/Documento_completo.pdf?sequence=1

Fombona Cadavieco, J., Rodríguez Pérez, C., y Barriada Fernández, C. (2012). Gestión de incidencias informáticas: el caso de la Universidad de Oviedo y la Facultad de Formación del Profesorado. *Revista de Universidad y Sociedad del Conocimiento*, *9*(2), 100-114. Recuperado de http://www.raco.cat/index.php/RUSC/article/viewFile/284627/372853

García, A., Corso, C., Gibellini, F., y Rapallini, M. (2013). Análisis de incidentes informáticos usando modelos de asociación y métodos del análisis de datos multivariante. *XV Workshop de Investigadores en Ciencias de la Computación* (pp.107-111). Universidad Tecnológica Nacional. Recuperado de http://sedici.unlp.edu.ar/bitstream/handle/10915/27107/Documento_completo.pdf?sequence=1

Gupta, R., Prasad, K. H., y Mohania, M. (2008). Automating ITSM Incident Management Process. *International Conference on Autonomic Computing* (pp. 141-150). Chicago: IEEE. doi:10.1109/icac.2008.22

Han J., y Kamber, M. (2006). *Data Mining: Concepts and Techniques*. (2.ª ed.). San Francisco: Morgan Kaufmann Publishers.

Han, J., Kamber, M., y Pei, J. (2012). *Data Mining. Concepts and Techniques*. (3.ª ed.). Boston: Morgan Kaufmann Publishers/Elsevier.

Sharma, S. & Bhatia, S. (2016). Analysis of association rule in data mining. *ICTS 16 Second International Conference on Information and Communication Technology for Competitive Strategies*, Udaipur, India, March 04-05, 2016. doi: 10.1145/2905055.2905238