Modelo de clasificación de incidentes tecnológicos para una empresa aseguradora desde un enfoque *machine learning*

Paola Andrea Gómez Jaramillo

Trabajo de investigación para optar por el título de Magíster en Ingeniería

Director

Favían González Echavarría, Magister en economía

Programa de Maestría en Ingeniería

Universidad de Antioquia

Facultad de Ingeniería

Maestría en Ingeniería

Medellín, Antioquia, Colombia

2022

Cita	(Gómez Jaramillo, 2022)
Referencia **Estilo APA 7 (2020)**	Gómez Jaramillo, P. A (2022). *Modelo de clasificación de incidentes tecnológicos para una empresa aseguradora desde un enfoque machine learning 2022* [Tesis de maestría]. Universidad de Antioquia, Medellín, Colombia.

Maestría en Ingeniería

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)
Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: Sara Cristina Vieira Agudelo.

Dedicatoria

Para quienes en vida me motivaron a perseverar en este propósito y ya no están aquí. Para mi familia y amigos que me apoyaron hasta llegar al final. Para mi esposo. Para quienes se alegran con mi alegría y para quienes celebran mis triunfos como propios. ¡Pero también para quienes dudaron que esto sería una realidad, lo logré!!!

Agradecimientos

Infinitas gracias a mis asesores (Jorge Perez y Favian González) porque confiaron en mí y me dieron regalos invaluables durante este acompañamiento hasta finalizar la elaboración de esta tesis de maestría, fue gracias a ustedes que logré mejorar mis habilidades en investigación, perseverar y llegar a publicar mi primer artículo científico. Deseo que cada estudiante de la Universidad que se interese por la investigación, no encuentre obstáculos en sus asesores para aprender, sino que encuentre facilitadores del aprendizaje como los que Dios me regaló.

Índice general

Índice de tablas

Índice de figuras

1 Introducción

Ante el aumento de los requerimientos en el uso de los servicios de las áreas de infraestructura tecnológica (TI) de las compañías, se han adoptado *softwares* en los procesos de gestión de incidentes (La ITIL define el incidente como una interrupción imprevista de un servicio informático o una reducción de la calidad de un servicio), que permiten realizar las principales actividades de entrada a dicho proceso de manera automática. Entre las actividades de mayor interés por los estudios revisados en el estado del arte se encuentran las relacionadas con la clasificación o asignación de incidentes, pues estas sugieren que la intervención de dicha actividad tiene un impacto directo en la mejora del desempeño del sistema de gestión de incidentes. Algunos estudios encontrados tienen un segundo propósito, este es, predecir fallas en equipos servidores a través del análisis de los incidentes y su correcta clasificación, teniendo en cuenta que la gestión de incidentes, según la Biblioteca de Infraestructura de Tecnología de la Información (ITIL), tiene como objetivo resolver cualquier problema que cause una interrupción en el servicio de manera rápida y eficaz. El alcance de este trabajo es la intervención de la actividad de clasificación de incidentes.

Con la información registrada actualmente en la herramienta tecnológica, utilizada por la mesa de ayuda (La ITIL define mesa de ayuda como un equipo de soporte que se encargada de capturar la demanda de resolución de incidentes y solicitudes de servicio relacionados con productos tecnológicos), se pretende automatizar la actividad de clasificación de incidentes de una compañía aseguradora en Colombia a través de métodos de aprendizaje supervisado (*machine learning*). El objetivo de intervención de dicha actividad es reducir la probabilidad del reproceso de la actividad de clasificación y aumentar la eficiencia en los tiempos de solución de incidentes. Es importante mencionar que la clasificación manual de incidentes es una actividad que requiere

de un análisis previo de datos que se capturan al momento de registrar el incidente. En la literatura revisada, se menciona que esta actividad manual es de riesgo, pues es posible realizar clasificaciones erradas a partir de análisis incorrectos de la información registrada por el cliente o durante la llamada de reporte de este.

La información capturada sobre los incidentes reportados depende de la necesidad específica de cada compañía, dado que esta define qué variables son relevantes para la toma de decisiones. La compañía aseguradora objeto de estudio genera los informes de gestión de los incidentes reportados al identificar los campos que considera relevantes al momento del registro del incidente. Entre las decisiones de interés en el presente trabajo se encuentra la asignación de incidentes a equipos de solución que hacen parte del área de tecnología y que pueden dar soporte y solución a los usuarios que encuentran inconvenientes con los aplicativos que intervienen en la operación. Se resalta el atributo de descripción del incidente, puesto que este contiene la mayor información sobre el reporte; mayormente se trata de textos libres, no estructurados.

Con el fin de cumplir el objetivo propuesto, se implementan los métodos de aprendizaje supervisado más utilizados en los estudios que se revisaron y otros de menor uso con el fin de obtener una comparativa incluyente. El proceso para seleccionar un modelo de aprendizaje automático personalizado, preciso y adecuado para la clasificación de incidentes de un problema específico, no es una tarea trivial; esto depende de varios factores: datos de entrenamiento, preprocesamiento de texto, vectorización de características, algoritmo de aprendizaje automático y parámetros de algoritmo (Al-Hawari y Hala, 2019). Según Altintas y Cuneyd (2014), el desempeño de la clasificación de un algoritmo varía directamente en relación con el algoritmo de aprendizaje automático, el método de ponderación y el conjunto de datos.

2 Planteamiento del problema

2.1 Descripción del contexto

La gestión de incidentes es un área de procesos perteneciente a la gestión de servicio de TI en las compañías. Su principal objetivo consiste en recuperar el nivel habitual del funcionamiento del servicio y minimizar el impacto negativo para que la calidad del servicio y la disponibilidad se mantengan (Ministerio de Tecnologías de la Información y las Comunicaciones, 2020). El proceso de gestión de incidentes es ejecutado por una mesa de ayuda que, a su vez, cuenta con un *software* que permite realizar la función de registro y canalización de incidentes para su solución. Estas mesas de ayuda cuentan con un nivel básico de resolución que se compone de agentes que reciben el reporte de falla y que conocen qué equipo de soporte debe direccionarlos en caso de que no puedan dar solución a quien reporta. Con el fin de realizar la clasificación de los incidentes, el personal de la mesa de ayuda realiza el registro del reporte y un análisis de este para proceder con la clasificación manual; esta debe realizarse por un técnico o especialista de mayor nivel o conocimiento para su resolución. Esto es lo que ITIL define como un escalado funcional de una mesa de ayuda. Generalmente se presentan los siguientes problemas en la gestión de incidentes al clasificar de manera manual un incidente al equipo de solución:

- Usuarios inconformes con el tiempo de respuesta y resolución de los incidentes reportados. Actualmente se cuenta con un recurso de apoyo del negocio para validar el cumplimiento de los niveles de servicio (cinco días hábiles) por quien reporta, pues se presenta un aumento en las quejas de usuarios con incidentes que superan los 10 días hábiles.

- Subutilización del recurso especializado al que se escalan los incidentes para su resolución. El área especializada en la resolución de estos indica que deben invertir un tiempo en el análisis de los incidentes que les escalan, pero no todos son de su alcance; por ello, deben reclasificarlos para que lleguen a los equipos de resolución correctos.
- Se contaba con algoritmos dentro de la herramienta tecnológica para realizar la clasificación de los incidentes, pero estos fueron cancelados; así, se definió su paso a una clasificación manual realizada por personal externo a la compañía. En ese sentido, se realiza la validación con el área encargada, la cual indica que, efectivamente, esta función fue entregada al proveedor contratado.
- El personal que actualmente realiza la clasificación en la compañía presenta altos índices de rotación, por lo que no cuentan con la curva de aprendizaje necesaria para hacer una correcta clasificación de los incidentes.

2.2 Definición del problema

La automatización de las actividades que componen el proceso de gestión de incidentes de TI se ha convertido en una necesidad vital para las compañías que dependen en gran medida de los servicios y recursos tecnológicos, pues el volumen de incidentes aumenta significativamente a medida que se incluyen en la operación herramientas tecnológicas. La automatización de las tareas cotidianas de TI (por ejemplo: la asignación de incidentes a los agentes de servicio, las notificaciones por correo electrónico para partes relacionadas, etc.) puede ayudar a definir y minimizar las actividades que comprenden los diferentes procesos empresariales. Además, permite evaluar el desempeño general de las áreas de TI sobre la base de los informes generados y los indicadores clave de desempeño evaluados (Al-Hawari y Hala, 2019).

En los últimos años, los proveedores de servicios de tecnologías de la información han tenido que reducir el costo de la prestación de servicios, mejorar la calidad y rapidez en la entrega de sus servicios. Aunque la calidad de la prestación de servicios de TI es multidimensional, la medida clave es claramente la disponibilidad de los aplicativos, especialmente para los de producción críticos. En consecuencia, debido a su impacto directo en la disponibilidad del sistema, la gestión de incidentes ha recibido mucha atención en cuanto a la búsqueda de una mayor calidad, resoluciones más rápidas y menores tiempos de inactividad. Las innovaciones recientes en este campo incluyen la capacidad adicional de clasificar los *tickets* (incidentes tecnológicos) para un enrutamiento óptimo a los equipos de resolución; la obtención de información pertinente del sistema y los *tickets* de incidentes similares para un mejor y más rápido análisis de causas raíz; y la dotación óptima de personal de los equipos de resolución.

El problema de clasificación de incidentes, según Maksai et al. (2014), es un problema multiclase que tiene como objetivo asignar la etiqueta correcta a cada incidente basado en las características de entrada. Desde el comienzo, la asignación correcta de un incidente al equipo de resolución correcto –es decir, mientras se abre el incidente– es fundamental para dirigirlo rápidamente y minimizar su tiempo de resolución (Al-Hawari y Hala, 2019). Asimismo, los autores indicaron que una mala clasificación impacta negativamente en el uso eficiente del recurso humano, el tiempo de solución y la satisfacción de los usuarios; por lo tanto, es necesario desarrollar y utilizar un modelo de clasificación de entradas basado en el aprendizaje automático. El uso de técnicas de aprendizaje automático ofrece mayor flexibilidad y precisión que una simple búsqueda de palabras clave. En los últimos años se ha avanzado mucho en el campo del aprendizaje automático y en la práctica se están utilizando técnicas de recuperación automática o semiautomática de la información en ámbitos muy diferentes (Sulaman, Weyns, & Host, 2015).

Asociar un incidente con una etiqueta de varias posibles es básicamente un problema de clasificación multiclase; así, dicho problema se resuelve al usar algoritmos supervisados, pues estos aprenden un modelo de clasificación basado en datos de entrenamiento de entrada de muestras (incidentes preetiquetados con el servicio respectivo correcto).

2.3 Objetivos

Con el fin de implementar una solución de automatización para el problema expuesto en una empresa aseguradora desde un enfoque *machine learning*, se plantearon los siguientes objetivos:

2.3.1 Objetivo general.

Desarrollar un modelo de clasificación de requerimientos para la suscripción de pólizas de seguros por nivel de escalamiento bajo un enfoque *machine learning*.

2.3.2 Objetivos específicos.

Evaluar qué método, entre árboles de clasificación, regresión logística, máquinas de soporte vectorial, *random forest* y Análisis Discriminante Lineal (ADL), presenta las mejores tasas de clasificación de requerimientos para la suscripción de pólizas de seguros por nivel de escalamiento.

Determinar si un modelo basado en el método *machine learning*, con las mejores tasas de clasificación de requerimientos para la suscripción de pólizas de seguros por nivel de escalamiento, es capaz de clasificar nuevas instancias de forma más precisa y eficiente, en comparación con el modelo basado en la interpretación del analista.

2.4 Preguntas de investigación

El presente estudio busca aprovechar el problema práctico descrito para una compañía de seguros de Colombia, un laboratorio de observación en dicho sector económico no reportado por la literatura revisada hasta el momento, en lo que respecta a la clasificación de incidentes tecnológicos desde un enfoque *machine learning*. Esto, en cuanto a la realización de investigaciones futuras y la ampliación de la comprensión del abanico de sectores para aplicar la clasificación de incidentes desde el mencionado enfoque (Paramesh y Shreedhara, 2019; Akbar y Jianglei, 2018; Maksai et al., 2014; Silva et al., 2018).

Mediante la observación de dicho sector, se busca profundizar en el tema e inducir nuevos hallazgos sobre la utilidad del *machine learning* y sobre las condiciones de uso que favorecen dicha utilidad en el sector de las compañías aseguradoras. Además, se realizarían comparaciones entre los diferentes métodos de aprendizaje con el fin de determinar cuál es el de mejor capacidad predictiva en el sector asegurador. Así, la presente propuesta de tesis de maestría en ingeniería, propuso enfocarse en dos preguntas concretas:

- P.1 ¿Qué método, entre árboles de clasificación, máquinas de soporte vectorial, *random forest*, regresión logística y LDA, presenta las mejores tasas de clasificación de incidentes para la suscripción de pólizas de seguros?

  Respecto a esta primera pregunta, se evidencia la necesidad de profundizar en cada uno de los métodos descritos y operativizarlos a través de algoritmos específicos para el sector objeto de estudio; ello, al considerar las restricciones impuestas (disponibilidad de datos, atributos aprovechables, lenguaje, narrativas, metadatos, etc.). Adicionalmente, es necesario diseñar y proveer una comparativa fiable y válida para concluir sobre el método

con mejor desempeño en el contexto de prueba. Ahora bien, al obtener respuestas a la primera pregunta, es necesario también ahondar en el tema a través de la segunda pregunta.

- P.2 ¿Un modelo entrenado bajo el método de *machine learning*, que presentó las mejores tasas de clasificación de incidentes para la suscripción de pólizas de seguros, es capaz de clasificar nuevas instancias de forma más precisa y eficiente en comparación con el modelo basado en la interpretación del analista?

  Esta segunda pregunta exige una comparativa adicional entre lo humano y lo artificial, en lo que respecta al problema de prueba. Por consiguiente, en esta pregunta también subyacen implicaciones prácticas para la empresa aseguradora.

2.5 Justificación e importancia

La actividad de entrada al proceso de gestión de incidentes es la de clasificación, según Jan et al. (2013). Cuando esta es realizada por humanos, se pueden presentar casos de registro donde falta contenido o se incurre en incoherencias en la información; en consecuencia, se dan errores en la clasificación de los incidentes reportados, lo que puede generar un aumento en los tiempos de respuesta, en tanto que se realizan las correcciones pertinentes. Asimismo, Li y Zhan (2012) y Yosifova et al. (2020) afirmaron que, cuando los tiempos de resolución de incidentes exceden los acuerdos de servicio, el incidente debe ser escalado a otros niveles de apoyo funcional apropiados para resolverlo. Sin embargo, el hecho de que el operador que realiza la actividad manual de clasificación tenga conocimientos en el ámbito pertinente, también se constituye como un factor importante que influye en la exactitud de la etiqueta asignada.

Se evidencia en la literatura un interés por automatizar esta actividad. Según indicaron Agarwal et al. (2012), una mejor asignación y un uso eficaz de los recursos de apoyo en organizaciones grandes se traduce en reducciones sustanciales de los costos (Gupta et al., 2008).

Si es posible la clasificación automática, esta puede acelerar el proceso de enrutamiento de incidentes al sugerir etiquetas apropiadas para los requerimientos entrantes (Son et al., 2014).

La calidad de la prestación de servicios de TI depende de múltiples dimensiones, pero la medida clave es la disponibilidad del sistema. En consecuencia, la gestión de incidentes se ha enfocado cada vez más en dar soluciones en el menor tiempo posible y minimizar los tiempos de inactividad. Las innovaciones recientes incluyen la capacidad adicional de clasificar los incidentes para lograr un enrutamiento óptimo a los equipos de resolución; la mejora de la calidad de la información pertinente para el sistema y de incidentes similares para un mejor y más rápido análisis de causas raíz; y, finalmente, la dotación óptima de personal de los equipos que se encargan de la resolución (Giurgiu et al., 2017).

Muchas empresas utilizan un sistema de seguimiento de incidentes (ITS), donde capturan y realizan su respectivo seguimiento. Los requerimientos pueden llegar por diferentes vías: redes sociales, correos electrónicos, aplicaciones móviles, llamadas telefónicas, sistemas de monitoreo y formas web (Marcu et al., 2009). Según Maksai et al. (2014), los textos de las entradas de los incidentes reportados son diferentes en todos los entornos de TI de las diferentes organizaciones, dado que están escritos por equipos que, a su vez, utilizan diferentes sistemas de monitorización y gestión de incidentes; ello hace que la reutilización de entradas etiquetadas manualmente y la transferencia de conocimiento entre diferentes entornos de TI sean inviables. Los incidentes, de acuerdo con la información capturada, se caracterizan y se almacenan en el sistema de gestión utilizado, definido por cada organización. Las características de este pueden ser continuas, categóricas o binarias.

Cada empresa determina la manera en que registra la información del incidente y las categorías que son relevantes para realizar sus análisis. En los estudios revisados, es posible

encontrar que la clasificación se da principalmente al utilizar el título o asunto y la descripción del incidente; como se indicaba, esta información se encuentra generalmente en textos libres y no estructurados. Esto también puede ser un inconveniente, debido a que, según Li y Zhan (2012), los usuarios y operadores son diversos en conocimientos previos y hábitos de escritura, lo que igualmente puede dar lugar a una descripción inadecuada de la falla reportada y, por tanto, la asignación al equipo puede ser incorrecta.

Según la ITIL, dentro del proceso de gestión de incidentes, la actividad de clasificación de estos tiene como fin establecer su impacto en la organización y su prioridad de resolución, dependiendo de su urgencia y su impacto, se asignan unos recursos y se establece un tiempo de resolución. Este tiempo, su impacto y su urgencia pueden variar a lo largo del análisis de la incidencia: pueden ampliarse por fallos en la estimación o recortarse por soluciones temporales eficaces para el cierre de la incidencia. Según Paramesh y Shreedhara (2019), la selección manual de cualquiera de las categorías de información del *ticket* puede resultar en una elección incorrecta debido al desconocimiento ante el incidente o a un error humano. Asimismo, Silva et al. (2018) afirmaron que, cuando se realiza una clasificación incorrecta de incidentes al asignarlos a grupos de resolución que no son capaces de resolverlos, se impide que el proceso de gestión se ejecute dentro de los tiempos planeados y de acuerdo con los niveles de servicio. Ello, a su vez, genera insatisfacción en los clientes. Automatizar el proceso de clasificación de incidentes significa evitar el error humano, reducir el desperdicio de recursos y evitar el enrutamiento incorrecto a causa de una clasificación errónea. Se encontró que el 95 % de los estudios revisados buscan solucionar el problema de la clasificación manual a través de la automatización de la clasificación de incidentes a través del aprendizaje supervisado; por mencionar algunos: Silva et al. (2018), Maksai et al.

(2014), Son et al. (2014), Akbar y Jianglei, (2018), Nguyen et al. (2016), Jan et al. (2013), Kallis et al. (2019), Gore et al. (2018) y Li y Zhan (2012).

A menudo, las personas son propensas a cometer errores durante los análisis; posiblemente porque intentan establecer relaciones entre múltiples características. Esto hace que les resulte difícil encontrar soluciones a problemas específicos. Generalmente, el aprendizaje automático puede aplicarse con éxito a estos problemas, con lo que mejora la eficiencia de los sistemas y el diseño de las máquinas. Cada instancia, en cualquier conjunto de datos utilizado por algoritmos de aprendizaje automático, se representa con el mismo grupo de características; estas pueden ser continuas, categóricas o binarias. Si se dan ejemplos con etiquetas conocidas, es decir, con los resultados correctos correspondientes, entonces el aprendizaje se llama supervisado; en contraste, si el aprendizaje no es supervisado, las instancias no están marcadas.

3 Marco conceptual

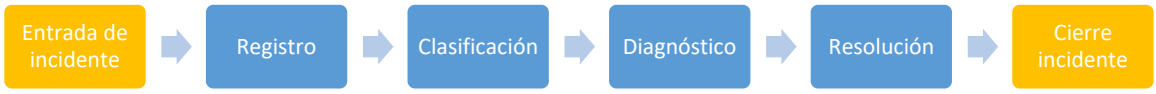
3.1 Definición de incidente

Con el uso exponencial de las tecnologías de la información en las empresas, se generan grandes volúmenes de incidentes asociados a la disponibilidad y el funcionamiento de la TI que soporta la operación. Estos incidentes deben detectarse y resolverse lo antes posible para no alterar la continuidad del servicio. La ITIL definió el incidente como una interrupción imprevista de un servicio informático o una reducción de la calidad de un servicio. Las compañías de servicios tecnológicos son las encargadas de mantener y gestionar mesas de ayuda, *hardware*, *software*, seguridad informática, etc. (Silva et al., 2018).

3.2 Proceso de gestión de incidentes

El proceso de gestión de incidentes es ejecutado por una o varias mesas de ayuda, dependiendo del tamaño y complejidad de la empresa. Altintas y Cuneyd (2014), afirmaron que, al tratarse de un proceso totalmente virtual, es decir, que no existe interacción personal entre el usuario final y un encargado de recibir dichos requerimientos, las empresas han intervenido con el propósito de dar una solución oportuna a la dificultad del cliente. En muchos casos, este proceso no es totalmente sistemático, y puede ser incoherente e ineficiente; debido a esto, se ha incrementado el interés en adoptar herramientas para ayudar y apoyar a los equipos responsables del proceso de gestión de incidentes a mejorar la eficiencia a través de herramientas como las ITS (Silva et al., 2018). El proceso de gestión de incidentes, según la ITIL, se presenta en la Figura 1, donde se evidencian las siguientes actividades:

Figura 1. Tomado de la ITIL



Fuente: elaboración propia

Por su parte, Silva et al. (2018) describieron las características principales de las actividades del proceso de gestión de incidentes diagramado, las cuales buscan dar respuesta oportuna a las fallas reportadas por los usuarios afectados. El proceso comienza con la actividad de registro, cuyo objetivo es que el incidente se registre tan pronto como haya sido detectado; si es posible, antes de causar impactos a otros usuarios. Luego, se tiene la clasificación, que implica asignar el incidente al equipo que puede darle solución, así como la designación de la prioridad. La actividad de diagnóstico incluye un análisis inicial para encontrar la solución: si esta se encuentra, se resuelve el incidente; en caso contrario, se escala al equipo de soporte pertinente.

Seguidamente, se desarrolla la etapa de resolución, donde se aplica la solución; esta debe ser comprobada para garantizar que el sistema está operativo. Por último, se tiene el cierre; este implica que el incidente ha sido asignado correctamente al equipo para su solución y, además, se debe validar la satisfacción del usuario con la solución y la documentación del incidente.

Como se indicó, la actividad de entrada al proceso de gestión de incidentes es la de clasificación. Según Jan et al. (2013), cuando dicha actividad es realizada por humanos, se pueden presentar casos de registro donde falta contenido o se incurre en incoherencias en la información; en consecuencia, se dan errores en la clasificación de los incidentes reportados, lo que puede generar un aumento en los tiempos de respuesta, en tanto que se realizan las correcciones pertinentes. Asimismo, Li y Zhan (2012) y Paramesh y Shreedhara (2019) afirmaron que, cuando los tiempos de resolución de incidentes exceden los acuerdos de servicio, el incidente debe ser escalado a otros niveles de apoyo funcional apropiados para resolverlo. Sin embargo, el hecho de que el operador que realiza la actividad manual de clasificación tenga conocimientos en el ámbito pertinente también se constituye como un factor importante que influye en la exactitud de la etiqueta asignada.

Se evidencia en la literatura un interés por automatizar esta actividad. Según indicaron Son et al. (2014), una mejor asignación y un uso eficaz de los recursos de apoyo en organizaciones grandes se traduce directamente en reducciones sustanciales de los costos. Si es posible la clasificación automática, esta puede acelerar el proceso de enrutamiento de incidentes al sugerir etiquetas apropiadas para los requerimientos entrantes.

La calidad de la prestación de servicios de TI depende de múltiples dimensiones, pero la medida clave es claramente la disponibilidad del sistema. En consecuencia, la gestión de incidentes se ha enfocado cada vez más en dar soluciones en el menor tiempo posible y menores tiempos de

inactividad. Las innovaciones recientes en este campo incluyen la capacidad adicional de clasificar los incidentes para lograr un enrutamiento óptimo a los equipos de resolución, la mejora de la calidad de la información pertinente del sistema y de incidentes similares para un mejor y más rápido análisis de causas raíz, y finalmente la dotación óptima de personal de los equipos que se encargan de la resolución (Giurgiu et al., 2017).

3.3 Características de un incidente

Los incidentes para el reporte de falla pueden llegar por diferentes vías: redes sociales, correos electrónicos, aplicaciones móviles, llamadas telefónicas, sistemas de monitoreo y formas web (Altintas y Cuneyd, 2014). Para ello, los sistemas de seguimiento de incidentes permiten realizar la captura de los datos relevantes, a fin de identificar la falla reportada y realizar la gestión y la trazabilidad hasta el cierre. Según Maksai et al. (2014), la información que ingresa en forma de textos estructurados y no estructurados es diferente en todos los entornos de TI de las diferentes organizaciones, pues estos están escritos por equipos que, a su vez, utilizan diferentes sistemas de monitorización y gestión de incidentes; ello hace que la reutilización de entradas etiquetadas manualmente y la transferencia de conocimiento entre diferentes entornos de TI sean inviables.

Al realizar el reporte del incidente, el usuario (cliente u operador) registra el incidente y almacena la información sobre este de acuerdo con lo definido para su escalamiento y resolución. Según Paramesh y Shreedhara (2019), algunos campos comunes que caracterizan un incidente son los siguientes:

- Categoría: etiqueta a través de la cual se asigna un equipo encargado de la solución del incidente. Ejemplo: seguridad, desarrollo, ataques cibernéticos.
- Prioridad: asignación de prioridad. Ejemplo: alto, medio, bajo.

- Descripción: describe la situación de falla. Ejemplo: qué, cómo, en qué aplicativo ocurre la falla, qué operación afecta, etc.
- Remitente: se trata del usuario o el cargo de la organización que es afectado por la falla reportada.
- Estado: se trata de una asignación que se da al incidente, depende del nivel de progreso hasta llegar a la solución. Ejemplo: abierto, suspendido, cerrado, etc.
- Nombre de quien resuelve: equipo encargado o persona encargada de la solución.
- Comentario de resolución: comentario con el que se cierra el incidente, generalmente indica lo que se hizo para resolverlo y si hubo comunicación con el usuario para validar que efectivamente fue solucionado.

Cada empresa determina la manera en que registra la información del incidente y las categorías que son relevantes para realizar sus análisis. En los estudios revisados, es posible encontrar que la clasificación se da principalmente al utilizar el título o asunto y la descripción del incidente; como se indicaba, esta información se encuentra generalmente en textos libres y no estructurados. Esto también puede ser un inconveniente, debido a que, según Li y Zhan (2012), los usuarios y operadores son diversos en conocimientos previos y hábitos de escritura, lo que igualmente puede dar lugar a una descripción inadecuada de la falla reportada y, por tanto, la asignación al equipo puede ser incorrecta.

3.4 Machine learning

La definición que presenta IBM es, Machine learning es una forma de la Inteligencia Artificial (AI) que permite a un sistema aprender de los datos, en lugar de aprender mediante la programación explícita. En cuanto a la literatura revisada, se encontró que el problema de la

clasificación de incidentes tecnológicos ha sido abordado bajo el enfoque *machine learning*; esto, dado que se trata de una tarea de aprendizaje supervisado, donde el objetivo es que la máquina aprenda a través de etiquetas preclasificadas y que se asigne automáticamente una etiqueta. Existen dos enfoques de clasificación de textos: binaria y multiclase (Silva et al., 2018). La literatura encontrada sobre incidentes tecnológicos aborda el problema de clasificación en lo concerniente a los textos; esto tiene dos objetivos: la asignación correcta de elementos al conjunto correspondiente de acuerdo con las condiciones definidas por el investigador o por expertos, y la predicción de fallas o errores en equipos o en procesos (Gore et al., 2018).

La clasificación de textos es el estudio de un conjunto de documentos textuales que pueden encontrarse en lenguaje natural o no, cuya respectiva asignación se da en categorías predefinidas (Gore et al., 2018). Para realizar tal clasificación, existe una gran variedad de métodos; entre estos, *random forest*, LDA, árboles de clasificación, regresión logística y SVM. Estos son utilizados en la literatura revisada con el fin de obtener una comparativa entre sus respectivos resultados de precisión.

3.5 Uso de los términos *Agrupación*/*clasificación*

Es común en la literatura encontrar el uso de los términos agrupación y clasificación. Con el fin de clarificar cada término, y con base en el enfoque *machine learning*, a lo largo de esta propuesta se emplea el término *clasificación* para hablar de la asignación que se realiza sin definición previa de etiquetas. Este tipo de entrenamiento se define como el aprendizaje no supervisado. Mientras que el término agrupación implica la definición previa de categorías para asignar cada elemento entrante, este tipo de entrenamiento se conoce como aprendizaje supervisado. Un algoritmo de clasificación tiene como objetivo crear un modelo, que representa

la relación entre los valores de los atributos del predictor y los valores de clase (etiquetas). (Otero, Freitas, & Johnson, 2012).

3.6 Ingeniería de características

Es el proceso de tomar un conjunto de datos y construir variables explicativas –características– que se pueden usar para entrenar un modelo de aprendizaje automático para un problema de clasificación/predicción. Dicho proceso se compone de dos actividades (Microsoft, 2020):

- Diseño de características: el proceso de crear nuevas características a partir de datos sin procesar para aumentar la eficacia predictiva del algoritmo de aprendizaje. La ingeniería de características debe capturar información adicional que no se pueda obviar fácilmente en el conjunto de características originales.
- Selección de características: el proceso de seleccionar el subconjunto de claves de las características en un intento por reducir la dimensionalidad del problema de entrenamiento.

La ingeniería de características se aplica previamente a la creación del modelo de clasificación en el que se hace un análisis, una limpieza y una estructuración de los campos de los datos. Este proceso es uno de los más importantes y más costosos del proceso de clasificación/predicción. El objetivo es eliminar los campos que no aportan a tal proceso y organizarlos adecuadamente para que el modelo no reciba información que no le es útil o que podría arrojar clasificaciones/predicciones de baja confianza.

Los datos de entrenamiento se componen de una matriz formada por filas y columnas; cada fila de la matriz es una observación o un registro. Las columnas de cada fila son las características que describen cada registro, y las características especificadas en el diseño experimental deben caracterizar los patrones de los datos. Asimismo, las características diseñadas que mejoran el

entrenamiento proporcionan información que ayuda a diferenciar de mejor manera los patrones de los datos; pero este proceso es, en cierto modo, un arte. Las decisiones acertadas y productivas a menudo requieren conocimiento especializado.

4 Estado del conocimiento sobre el problema de clasificación de incidentes tecnológicos desde un enfoque de aprendizaje automático

Con el fin de conocer la manera en que diferentes autores han tratado el tema de estudio presentado, se realizó la búsqueda, la lectura y el análisis de la bibliografía relacionada. A continuación, se intenta establecer lo que se ha hecho recientemente sobre el tema de estudio a través de la extracción de lecciones aprendidas en formato tabular, donde se indica, para cada estudio, las características relevantes: dominio, objetivo, atributos de entrenamiento, variable a predecir, etapas, tamaño de la muestra, métodos probados, métodos de validación, métodos de mejor desempeño y desafíos futuros.

Tabla 1. *Dominio y objetivo principal del estudio*

Dominio	Objetivo	Estudios	# Estudios	% de estudios
Incidentes reportados a través de mesas de ayuda, servicio infraestructura TI.	Automatizar la tarea de clasificación de incidentes a través del aprendizaje supervisado con el fin de mejorar el desempeño del sistema de gestión de incidentes	Paramesh & Shreedhara (2019); Altintas, et al (2014); Silva, et al (2018); Maksai, et al (2014); Son, et al (2014); Han y Akbari (2018); Al-Hawari y Barham (2019); Nguyen, et al (2016); Jan, et al (2013); Kallis,et al (2019); Gore, et al (2018); Li y Zhan (2012).	12	75%
Incidentes reportados a través de mesas de ayuda, servicio infraestructura TI.	Automatizar la tarea de clasificación de incidentes a través del aprendizaje supervisado con el fin de predecir daños en servidores	Bogojeska, et al (2014); Giurgiu, et al (2017)	2	13%
Incidentes reportados a través de mesas de ayuda, servicio infraestructura TI.	Automatizar la tarea de clasificación de incidentes a través del aprendizaje supervisado con el fin de generar una matriz de riesgo que permita montorear el tema de salud ocupacional	Kurian, et al (2020)	1	6%
Incidentes reportados a través de mesas de ayuda, servicio infraestructura TI.	Desarrollar un prototipo de solución de un sistema que identifica automáticamente la información relativa a incidentes, a partir de textos disponibles en la web con el fin de mejorar el proceso de análisis de riesgos de una compañía de aviación. (Esto complementa el análisis de los incidentes reportados a la mesas de ayuda)	Sulaman (2015)	1	6%

Fuente: elaboración propia

La evidencia encontrada en la Tabla 1 indica que la automatización de las actividades que componen los procesos de los servicios de TI ha sido un tema de gran interés en los últimos años para la comunidad científica, específicamente la clasificación de incidentes en las mesas de ayuda de las empresas. En cuanto a los objetivos buscados en los estudios revisados, se logró identificar que el 75 % se enfoca únicamente en la automatización de la actividad de clasificación de incidentes; pero también se identificaron algunos estudios que presentan interés en extender este objetivo al análisis de riesgos. También se encontró un estudio que se interesa por elaborar una matriz de riesgos para monitorear el tema de salud ocupacional, dos estudios que buscan intervenir los servidores de manera proactiva y un estudio que busca incluir otros medios de posible identificación de riesgos, como las noticias en la web, para complementar lo que la información

que se captura de los incidentes no satisface por sí sola. Los estudios revisados son diversos en aspectos como la selección de los atributos de entrada para el modelo de clasificación en sus aplicaciones:

Tabla 2. *Atributo del modelo*

Atributo	Estudio	# Estudios
Descripción	Paramesh & Shreedhara (2019); Altintas, et al (2014); Sara Silva, et al (2018); Bogojeska, et al (2014); Giurgiu, et al (2017); Al-Hawari y Hala Barham (2019); Gore, et al (2018); Jan, et al (2013); Sulaman (2015); Li y Zhan (2012); Kallis,et al (2019).	11
Equipo de solución	Son, et al (2014); Nguyen, et al (2016).	2
Todos los atributos	Silva, et al (2018)	1
Texto de resolución	Maksai, et al (2014); Giurgi, et al (2017)	2
Fecha de reporte	Giurgiu, et al (2017)	1
Título	Paramesh & Shreedhara (2019); Al-Hawari y Barham (2019); Han y Akbari (2018).	3

Fuente: elaboración propia

Tabla 3. *Variable a predecir*

Variable a clasificar	Estudio	# Estudios
Tipo de problema	Paramesh & Shreedhara (2019); Silva, et al (2018); Maksai, et al (2014); Son, et al (2014); Giurgiu, et al (2017); Gore, et al (2018); Kallis, et al (2019); Kurian, et al (2020)	8
Área de solución	Al-Hawari y Barham (2019); Altintas, et al (2014); Han y Akbari (2018).	3
Prioridad	Altintas, et al (2014); Nguyen, et al (2016)	2
Nombre servidor	Jan, et al (2013)	1
Tipo de solución	Nguyen, et al (2016)	1

Fuente: elaboración propia

Las tablas 2 y 3 muestran el uso de los atributos en las diferentes aplicaciones que componen el estado del arte. Puede evidenciarse que más del 50 % de los estudios utilizan la descripción o el cuerpo del incidente como atributos del modelo de clasificación; asimismo, el asunto o título también es utilizado en un 25 % de los estudios revisados. Otros atributos, como el sujeto que reporta, la fecha, la criticidad, la prioridad y el texto de solución o cierre, son utilizados en menor medida. Cabe resaltar que solo uno de los estudios revisados utiliza todos los atributos

del incidente para la clasificación automática. Son et al. (2014) recomendaron el uso del asunto o título y de la descripción del *ticket*; ello, para obtener una mayor precisión del clasificador utilizado. Finalmente, existe una convergencia en que las variables a utilizar son de tipo categórico en todos los estudios.

Cada empresa determina la manera en que registra la información del incidente y las categorías que son relevantes para realizar sus análisis. En los estudios revisados, es posible encontrar que la clasificación se da principalmente al utilizar el título o asunto y la descripción del incidente; como se indicaba, esta información se encuentra generalmente en textos libres y no estructurados. Esto también puede ser un inconveniente, debido a que, según Li y Zhan (2012), los usuarios y operadores son diversos en conocimientos previos y hábitos de escritura, lo que igualmente puede dar lugar a una descripción inadecuada de la falla reportada y, por tanto, la asignación al equipo puede ser incorrecta.

En cuanto a la literatura revisada, se encontró que el problema de la clasificación de incidentes tecnológicos ha sido abordado bajo el enfoque *machine learning*; esto, dado que se trata de una tarea de aprendizaje supervisado, donde el objetivo es que la máquina aprenda a través de etiquetas preclasificadas y que se asigne automáticamente una etiqueta. Existen dos enfoques de condiciones definidas por el investigador o por expertos, y la predicción de fallas o errores en equipos clasificación de textos: binaria y multiclase (Silva et al., 2018). La literatura encontrada sobre incidentes tecnológicos aborda el problema de clasificación en lo concerniente a los textos; esto con el objetivo de realizar la clasificación de los documentos textuales en categorías predefinidas (Gore et al., 2018).

Con respecto a la metodología de investigación, los autores estudiados coincidieron en las siguientes etapas: recolección de la información, preprocesamiento de los datos, construcción del vector de características, selección del método de clasificación, entrenamiento del modelo, clasificación y validación comparativa entre métodos seleccionados. Durante la etapa de preprocesamiento de los datos, Silva et al. (2018) propusieron iniciar con la "tokenización" de texto, es decir, dividir los atributos seleccionados en todas las palabras que lo componen. Luego, con el diccionario resultante compuesto por todas las palabras diferentes presentes en las descripciones, se aplica la eliminación de términos no significativos para la clasificación; para esto se utiliza la frecuencia inversa del documento (TF-IDF), la cual consiste en asignar a cada término un peso basado en la frecuencia del término en el documento.

Por su parte, Jan et al. (2013) indicaron que los incidentes consisten mayormente de textos libres, no estructurados, que pueden incluir caracteres especiales, espacios, prefijos y sufijos, los cuales no deben considerarse; así, se pueden incluir reglas heurísticas para eliminarlas. Nguyen et al. (2016) afirmaron que un modelo de aprendizaje automático personalizado adecuado para la clasificación del servicio de *tickets* depende de factores como los datos de entrenamiento, preprocesamiento de texto, vectorización de características, algoritmo de aprendizaje automático y parámetros de algoritmo.

Con respecto a la fase de construcción del vector de características, Paramesh y Shreedhara (2019) señalaron que, para implementar los modelos de clasificación, los datos que se han de utilizar en la fase de entrenamiento deben ser convertidos a representación vectorial. Para la construcción del vector de características, Altintas y Cuneyd (2014) sugirieron aplicar la ponderación boleeana y la frecuencia de palabras. En la fase de selección del algoritmo, se observa

que no existe un criterio claro para realizar dicha selección, excepto el resultado de la validación de estudios que componen su estado del arte.

Para la fase de entrenamiento, autores como Paramesh y Shreedhara (2019) recomendaron dividir los datos históricos de la primera fase metodológica así: 80 % para entrenar y 20 % para la fase de validación. Algo similar sugirieron Kurian et al. (2020) cuando indicaron que el 70 % debería ser para entrenar y el 30 % para validar; igualmente, (Jan et al. (2013) indicaron que el 90 % debía ser para entrenamiento y el 10 % para validar. Al final, los autores realizaron la validación y la comparación de los algoritmos de clasificación utilizados. Durante esta fase, Gore et al. (2018); Altintas y Cuneyd (2014); Paramesh y Shreedhara (2019); Silva et al. (2018); Maksai et al. (2014); Giurgiu et al. (2017); Nguyen et al. (2016); Jan et al. (2013); y Kallis et al. (2019) sugirieron la validación cruzada (CV), una metodología de evaluación ampliamente utilizada para los sistemas de clasificación de textos.

Finalmente, en la selección del algoritmo de clasificación se consideró la Tabla 4, donde se muestra el porcentaje de uso de los métodos de aprendizaje supervisado en los estudios revisados. Es importante indicar que un mismo autor puede utilizar más de un método; ello, con el fin de realizar comparaciones de desempeño de estos para el modelo planteado.

Tabla 4. *Métodos de clasificación de aprendizaje supervisado utilizados en los estudios revisados*

Método	% de utilización
Naive Bayes	56%
Maquinas de soporte vectorial (SMV)	56%
Regresión logística	31%
K vecinos mas cercanos (KNN)	31%
Árboles de clasificación	31%
Máquinas de aumento de gradiente (GBM)	13%
Random forest	6%
Adaboost	6%
Redes neuronales	6%
CRF	6%

Fuente: elaboración propia

Como se observa en la Tabla 4, los dos métodos más utilizados fueron SVM y Naive Bayes, ambos usados en el 56 % de estudios. Las SVM clasifican nuevas instancias de ensayo al determinar un hiperplano que separa dos grupos de clases (Paramesh y Shreedhara, 2019). Aunque obtuvieron resultados sobresalientes en cuanto a precisión, Gore et al. (2018) explicaron que existen algunas desventajas; entre estas, se requiere un conjunto de entrenamiento negativo adicional, tiempo computacional y memoria adicional. El Naive Bayes es un algoritmo de clasificación probabilística muy simple basado en el teorema de probalidad de Bayes. Este algoritmo se utiliza en problemas de clasificación de textos o documentos debido a su regla de independencia, es decir, problemas donde las palabras en un documento no están relacionadas entre sí (Paramesh y Shreedhara, 2019).

La regresión logística, según Paramesh y Shreedhara (2019), es básicamente un clasificador binario, pero puede generalizarse para problemas multiclase. Por otro lado, el método KNN encuentra los documentos k más similares basados en la métrica de distancia y luego asigna la etiqueta de mayor clase al nuevo documento. Igualmente, utiliza la distancia euclidiana entre dos puntos como una métrica de distancia para encontrar los documentos k similares. De acuerdo con Maksai et al. (2014), el algoritmo GMB agrupa las entradas en conjuntos con textos similares y luego elige aleatoriamente un número de muestras de cada grupo; así, se seleccionan más muestras de los grupos menos homogéneos. En cuanto a lo que indicaron ambos autores, este método es potente y flexible y puede capturar eficazmente complejas dependencias de funciones no lineales, así como ofrecer resultados de alta calidad en términos de precisión, predicción y capacidad de generalización.

Random forest, como indicaron Giurgiu et al. (2017), es un método que consiste en un conjunto de árboles de decisión; estos consideran interacciones no lineales entre características de entrada. Este método no es tan popular como el SVM y el Naive Bayes, pero se basa en el principio de los árboles de clasificación, que sí es ampliamente utilizado en el 31 % de los estudios. Otros métodos, como las redes neuronales, no son muy explorados en los estudios revisados, dado que la naturaleza no estructurada de los datos utilizados no se presta para desarrollar las métricas necesarias para el entrenamiento (Gore et al., 2018).

El proceso para seleccionar un modelo de aprendizaje automático personalizado, preciso y adecuado para la clasificación de incidentes no es una tarea trivial, puesto que depende de varios factores: datos de entrenamiento, preprocesamiento de texto, vectorización de características, algoritmo de aprendizaje automático y parámetros de algoritmo (Al-Hawari y Hala, 2019). Según Altintas y Cuneyd (2014), el desempeño de la clasificación de un algoritmo varía directamente en

relación con el algoritmo de aprendizaje automático, el método de ponderación y el conjunto de datos. Finalmente, para realizar la evaluación del desempeño de los métodos, los autores utilizaron mayormente métodos de CV. En la Tabla 5, se presenta dicha matriz.

Tabla 5. *Matriz de confusión*

	Predicción de falla	**Predicción de éxito**	**Error de modelo**
Falla	a	B	$\frac{b}{a+b}$
Éxito	c	D	$\frac{c}{c+d}$
Error de usuario	$\frac{c}{a+c}$	$\frac{b}{b+d}$	Error general: $\frac{(b+c)}{a+b+c+d}$

Fuente: (Berk, 2017)

La matriz de confusión está compuesta por una serie de indicadores de precisión, sensibilidad y especificidad; cada uno permite determinar el desempeño del método implementado para el entrenamiento del modelo.

4.1 Profundización sobre los métodos a comparar en cuanto a estructura, procedimiento y recomendaciones para su aplicación

Esta sección contiene una breve explicación de los métodos aplicados a la solución del problema planteado. Se trata de métodos estadísticos que permiten desarrollar modelos para predecir respuestas a observaciones futuras. La predicción es un superconjunto de pruebas de hipótesis y estimación que se basa esencialmente en un conjunto de variables predictoras, cuyos

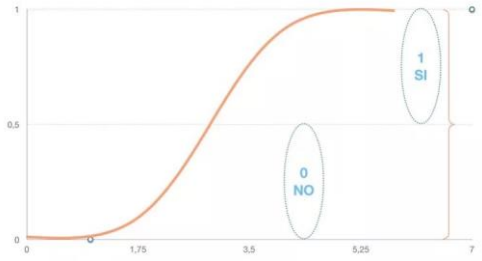
valores se espera que influyan en la asignación de etiquetas futuras; esto es lo que se llama "modelado predictivo". Los modelos predictivos a menudo se desarrollan al utilizar el aprendizaje supervisado, en el que, a partir de un conjunto de resultados predeterminados (etiquetados previamente), se identifican variables que pueden contribuir a predecirlos y se aplican algoritmos de análisis estadístico a un conjunto de datos de prueba para determinar qué variables son los predictores más relevantes y cómo deben ponderarse. La recolección de conjuntos de datos es un paso relevante en ese proceso.

Los modelos predictivos se han utilizado desde hace mucho tiempo en contextos de negocios para pronosticar desempeños financieros, modelar oferta y demanda, analizar el comportamiento de pago de préstamos, etc. En general, los modelos multivariados se usan ampliamente en áreas como medicina, epidemiología, investigación en salud, farmacéutica y otras afines, con aplicaciones sobre el diagnóstico y el pronóstico.

4.1.1 Regresión logística.

La regresión logística es uno de los algoritmos de *machine learning* más simples y más utilizados para la clasificación de dos clases, es fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria; esta describe y estima la relación entre una variable binaria dependiente y las variables independientes. Por otra parte, la regresión logística lleva el nombre de la función utilizada en el núcleo del método, la cual es también llamada función sigmoide. Esta es una curva en forma de *S* que puede tomar cualquier número de valor real y asignarle un valor entre 0 y 1 (Harrell, 2015).

Figura 2. Sigmoide regresión logística



Fuente: (Harrell, 2015)

Si la curva va a infinito positivo, la predicción se convierte en 1; si pasa el infinito negativo, la predicción se convierte en 0. Si la salida de la función sigmoide es mayor que 0,5, se puede clasificar el resultado como 1 o *sí*; y si es menor que 0,5, se puede clasificar como 0 o *no*.

Para denotar el modelo de regresión lineal, se tiene la variable dependiente o variable respuesta y X1, X2, X3, X4, ….. Xp; esta es una lista o vector de variables independientes o descriptoras de un sujeto –para este caso, incidente– de la población de interés. Suponiendo que β = β0, β1,..., βp denotan la lista de coeficientes de regresión (parámetros), β0 es un parámetro de intercepción opcional, y β1….. βp son pesos o coeficientes de regresión correspondientes a X1, …..Xp.

$$Y = \beta 0 + \beta 1X1 + \ldots + \beta pXp \text{ donde } X0 = 1$$

Por otra parte, la ecuación de la función sigmoide es la siguiente:

$$P = \frac{1}{1+e^{-y}} \quad \text{ó} \quad P = \frac{1}{(1+ e^{-(y = \beta 0 + \beta 1X1 + \ldots + \beta pX)})}$$

La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas, y en la construcción del modelo no hay ningún supuesto en cuanto a la

distribución de probabilidad de las variables; por ello, puede incluirse cualquier tipo de variable. El modelo de regresión logística puede considerarse como una fórmula para calcular la probabilidad de pertenencia a uno de los grupos, de manera que estima la probabilidad de que una observación pertenezca a uno de los grupos. La interpretación del resultado de la aplicación de esta metodología es sencilla por tratarse en términos de probabilidad. El modelo de regresión logística utiliza la estimación por máxima verosimilitud y estima la probabilidad de que un evento dado ocurra. Para construir un modelo de regresión logística se requiere:

- Un conjunto de variables independientes o predictoras.
- Una variable respuesta dicotómica o etiqueta que se pretenda asignar para la clasificación.

El coeficiente de determinación (R^2) que aparece en los modelos de regresión es una medida de la calidad del ajuste del modelo propuesto; se mide la proporción de variabilidad total de la variable dependiente respecto a su media.

4.1.2 Árboles de clasificación

Son una forma de regresión de estratificación con predictores que son variables indicativas. Estos implican la partición de los datos de acuerdo con el mejor resultado de suma de cuadrados que arroje. Existen muchos métodos estadísticos basados en árboles; entre ellos, los árboles de clasificación y regresión (CART) introducidos por Breiman et al. (1984), los cuales son aplicados cuando las variables de respuesta son categóricas y binarias. El algoritmo de CART se llama "codicioso", porque busca el mejor resultado sin mirar hacia atrás a las divisiones pasadas o hacia adelante a las divisiones futuras (Berk, 2017).

El CART construye particiones con una serie de límites rectos perpendiculares al eje del predictor utilizado; por esta razón, se consideran algoritmos fáciles de trabajar y con un buen desempeño en la práctica. Gráficamente, la partición del CART se muestra a menudo como un

árbol invertido; dentro de cada uno de los nodos terminales, se puede calcular la proporción de "éxitos" y la proporción de "fallos". Los efectos de interacción deben tenerse en cuenta cuando se interpretan los diagramas de árbol, y para determinar si las clases o etiquetas asignadas por CART fueron correctas, se utilizan las tablas de confusión, las cuales tabulan las clases observadas contras las clases que asigna el CART.

Si la tarea del árbol es clasificar, se puede calcular, dentro de cada uno de los nodos terminales, la proporción de "éxitos" y la proporción de "fallos". Asimismo, las proporciones condicionales se pueden utilizar para adjuntar etiquetas de clase a nodos terminales que, a su vez, se pueden asignar a observaciones. Las etiquetas de clase son un segundo tipo de valor ajustado; si la mayoría de las observaciones en un nodo terminal son de un tipo, todas las observaciones en esa partición pueden asignarse a una etiqueta. Las etiquetas son una buena suposición para el resultado binario desconocido para cada caso; a menudo, esto se desarrolla como pronóstico.

Con el fin de determinar qué tan correcta es la asignación de etiquetas del árbol de decisión, se puede realizar una evaluación a través de tablas de confusión que cruzan las clases observadas con las clases que el CART asigna. Se espera que tanto las clases clasificadas como las que arroja el algoritmo correspondan a la misma etiqueta. Una vez que se ha cultivado un árbol satisfactorio y se ha evaluado honestamente con los datos de la prueba, está listo para su uso; ello, cuando no se conoce la clase de resultado, pero los valores del predictor sí. Por lo general, hay cuatro tipos de evaluaciones del desempeño que se hacen a partir de cuadros de confusión.

- La proporción global de casos clasificados incorrectamente es una forma inicial de evaluar la calidad del rendimiento. Es simplemente el número de observaciones en las celdas fuera de la diagonal dividido por el número total de observaciones. Si todas las observaciones caen en la diagonal principal, CART, entonces se ha realizado perfectamente.

- El error general ignora que a menudo es más importante ser preciso para una de las clases variables de respuesta que para otra. Los dos tipos de fallas del modelo generalmente se llaman "falsos positivos" y "falsos negativos". Los éxitos incorrectamente llamados fracasos son falsos negativos, y los fracasos incorrectamente llamados éxitos son falsos positivos.
- Las proporciones de la columna abordan una cuestión algo diferente. Una de las condiciones de la clase ajustada es que esta calcula la proporción de veces que una clase ajustada es incorrecta, mientras que las proporciones de fila ayudan a evaluar qué tan bien ha funcionado el algoritmo CART. Las proporciones de columna ayudan a evaluar qué tan útiles son los resultados de CART si se ponen a trabajar.
- La relación entre el número de falsos negativos y el número de falsos positivos muestra cómo los resultados cambian un tipo de error por otro.

4.1.3 Random forest.

Es un método que surge como una mejora a los árboles de clasificación. Básicamente, es un método que combina una cantidad grande de árboles de clasificación independientes, probados sobre conjuntos de datos aleatorios con igual distribución. Se trata de un esquema propuesto por Breiman (2001), para construir un conjunto predictor con un conjunto de árboles de decisión que crecen en subespacios de datos seleccionados al azar. A pesar del creciente interés y uso práctico, ha habido poca exploración de las propiedades estadísticas de los bosques aleatorios, y poco se sabe acerca de las fuerzas matemáticas que impulsan el algoritmo. Breiman (1996) demostró que se pueden lograr mejoras sustanciales en la clasificación y la precisión de regresión mediante el uso de conjuntos de árboles, donde cada uno se cultiva de acuerdo con un parámetro aleatorio. Las predicciones finales se obtienen al agregarlas sobre el conjunto, así como los componentes básicos

del conjunto son predictores estructurados por árboles, y cada uno de estos árboles se construye al usar una inyección de aleatoriedad.

Random forest es un método rápido y fácil de implementar, produce predicciones muy precisas y puede manejar un gran número de variables de entrada sin necesidad de ajustes excesivos. De hecho, se considera que es una de las técnicas de aprendizaje de propósito general más precisas disponibles.

En el enfoque de Breiman et al. (1984), cada árbol de la colección se forma al seleccionar, primero al azar, en cada nodo, un pequeño grupo de coordenadas de entrada (también llamadas características o variables a continuación); ello, a fin de dividirse y, además, para calcular la mejor división basada en estas características en el conjunto de entrenamiento. El árbol se cultiva con el uso de la metodología CART (Breiman et al., 1984), a tamaño máximo, sin poda. Este esquema de aleatorización subespacial se combina con la idea de "ensacado" de Breiman (1996); Bühlmann y YU (2002); y Buja y Stuetzle (2006) para reconstruir, con reemplazo, el conjunto de datos de entrenamiento cada vez que se cultiva un nuevo árbol individual. El modelo está expresado de la siguiente forma:

Colección de árboles aleatorizados de regresión de base {rn(x,Θm,Dn),m 1}, donde Θ1,Θ2,... son salidas i.d. de una variable aleatoria Θ. Estos árboles aleatorios se combinan para formar la estimación de regresión agregada.

$$\overline{rn}\ (X, Dn) = \ \mathrm{E}\Theta\ [\mathrm{rn}(\mathrm{X}, \Theta, \mathrm{Dn}]$$

EΘ: expectativa con respecto al parámetro aleatorio, condicionalmente en X, y el conjunto de datos Dn.

Al tomar el promedio de los resultados individuales (procedimiento justificado por la ley de grandes números, ver el apéndice en Breiman (2001), la variable aleatorizadora Θ se utiliza

para determinar cómo se realizan los cortes sucesivos al construir los árboles individuales, como la selección de la coordenada a dividir y la posición de la división (Biau, 2012). El *random forest* utiliza el *bagging*, lo que implica que distintos árboles ven distintas porciones de los datos; así, ningún árbol ve todos los datos de entrenamiento. Esto hace que cada uno se entrene con distintas muestras para un mismo problema y, de esta forma, al combinar los resultados, unos errores se compensan con otros y se tiene una predicción que generaliza mejor. Dado que un *random forest* es un conjunto de árboles de decisión, y los árboles son modelos no paramétricos, este modelo tiene las mismas ventajas y desventajas de los modelos no paramétricos.

4.1.4 LDA.

Se trata de una técnica estadística multivariante, cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables medidas sobre estos. En caso de que existan, el método permite explicar en qué sentido se dan y facilitar procedimientos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos analizados. Se puede considerar este método como un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos, mientras que las variables independientes son continuas y determinan a qué grupos pertenecen los objetos. (de la Fuente, 2011, p. 3)

Para efectuar el análisis es necesario considerar una serie de supuestos:

- Se tiene una variable categórica y el resto de variables son de intervalo o de razón y son independientes respecto de ella.
- Se necesitan al menos dos grupos, y para cada grupo se necesitan dos o más casos.
- El número de variables discriminantes debe ser menor que el número de objetos menos 2, es decir, (x, x, x) 1 2 " p, donde p < (n−2), siendo n ≡ número de objetos.

- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.
- El número máximo de funciones discriminantes es el mínimo [número de variables, número de grupos menos 1] –con q grupos, (q−1), funciones discriminantes–.
- Las matrices de covarianzas dentro de cada grupo deben ser parecidas.
- Las variables continuas deben seguir una distribución normal multivariante. (de la Fuente, 2011, p. 3)

El modelo matemático

Partiendo de que existen q grupos en los que se asigna una serie de objetos y de p variables medidas sobre ellos (x_1, x_2, x_p), se intenta obtener para cada objeto una serie de puntuaciones que indican el grupo al que pertenecen (y_1, y_2, y_m), de modo que sean funciones lineales de (x_1, x_2, x_p):

$$y_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p + w_{10} \ldots\ldots$$

$$y_m = w_{m1}x_1 + w_{m2}x_2 + \cdots + w_{mp}x_p + w_{10}$$

Donde, m = mín[q−1,p].

Estas deben discriminar o separar lo máximo posible a los q grupos. Dichas combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos (Rolph et al., 2007). El criterio de clasificación de este método es:

Hipótesis: las distribuciones solo se diferencian por su localización (igual forma y varianza).

Se trata de minimizar los errores de clasificación.

- Si xi < C se clasifica en el grupo I.

- Si xi > C se clasifica en el grupo II.

El punto C se denomina punto de corte discriminante: $C = \frac{\overline{X_I} + \overline{X_{II}}}{2}$

4.1.5 SVM.

Las SVM o máquinas de vector soporte son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T. Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Las SVM tienen una fundamentación matemática y estadística pura dentro de la teoría estadística de aprendizaje; a pesar de esto, la implementación básica cuenta con algunas falencias, puesto que están diseñadas originalmente para problemas de clasificación binarios (dos clases). Además, hay un inconveniente: su algoritmo básico de entrenamiento genera gran cantidad de vectores soporte; ello ocasiona la lentitud y el sobreajuste en la clasificación. (Platzi, s.f., párr. 1)

Sea dado un conjunto de datos de entrenamiento $\{x_i, y_i\}$ con i = 1,..., l, $y_i \in \{-1, 1\}$ y xi $\in R^d$, existe un hiperplano que separa los datos de etiquetas positivas y negativas, así:

$$x_i w + b \geq 1 - \xi_i \quad para\ y_i = 1$$

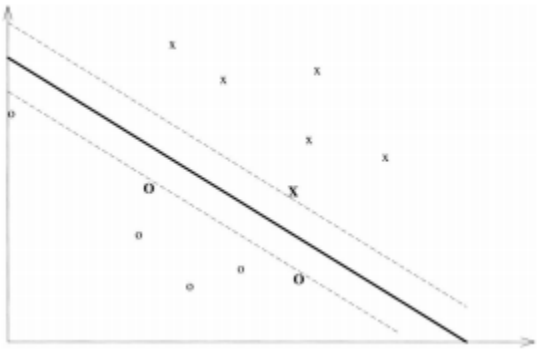
$$x_i w + b \geq 1 - \xi_i \quad para\ y_i = 1$$

$$\xi_i \geq \forall_i$$

En la fórmula anterior, w es la normal al hiperplano y ξi son las variables introducidas por los errores de clasificación en calidad de violaciones del hiperplano. Así, ∑ ξ es la cota del error de clasificación. Una manera directa de añadir el costo a la función objetivo es minimizar $\|w^2\| /2 + C \sum \xi i$, donde C es la constante elegida, correspondiente al inverso del valor de la penalización

de los errores (Gutierrez, 2007). De esa forma, se tiene un caso de optimización convexa como problema de optimización cuadrática cuya forma dual Wolfe está dada por:

Figura 3. Hiperplano de separación



Fuente: (Gutierrez, 2007)

En el diseño de las SVM, es importante tener en cuenta los siguientes elementos:

- Capacidad de generalización: dado que las SVM no minimizan una superficie de error, sino un margen que mide la separación entre las clases, el problema de optimización es convexo; por eso siempre tiene un mínimo global. Para esto, las SVM se orientan en la obtención de dicho mínimo mediante el algoritmo de la función dual de Wolfe.
- Funciones kernel y espacios no lineales: el uso del kernel en las SVM corresponde a una función de decisión que no es linealmente dependiente del espacio de entrada. Convenientemente, al reemplazar la función kernel en la SVM, se genera una máquina que existe en un espacio infinito dimensional y que se toma la misma cantidad de tiempo en la fase de entrenamiento que otras con datos no mapeados.

- Arquitectura: la arquitectura de las SVM solo depende del parámetro C, la función kernel (incluyendo sus parámetros). Para el caso del RBF, solo se requiere el parámetro σ; ello evita requerimientos sobre parámetros exclusivos de arquitectura, como el número de nodos y capas, el tipo de conexión entre capas, etc.
- Validación del error: en cuanto a la validación del error, las dos formas más utilizadas son: a) la prueba simple independiente (IDT), que consiste en partir el espacio de entrada en dos, uno para la fase de entrenamiento y otro para la de validación, con el fin de medir el desempeño del clasificador en cuanto a la generalización; y b) la CV, cuando varias particiones sirven a la vez de entrenamiento y validación.
- Algoritmos de entrenamiento y optimización: entre los algoritmos de optimización y entrenamiento más conocidos están el SMO (Secuential Minimal Optimization), propuesto por Bernhard Schölkopf. Este es el más utilizado debido a su efectividad computacional; descompone el problema de optimización en tareas mucho más pequeñas, con el fin de reducir el tamaño de las operaciones matriciales, es decir, los requerimientos de memoria y procesador.

En la Tabla 6 se presenta un resumen de las generalidades de los métodos supervisados de clasificación seleccionados para el entrenamiento del modelo planteado:

Tabla 6. *Resumen métodos de entrenamiento supervisado*

Método	**Modelización de la variable respuesta**	**Principio de funcionamiento**	**Modelo estadístico**
Regresión logística	X= variables predictoras Y= variable respuesta $Y = \beta 0 + \beta 1X1 + \ldots + \beta pX$ Y…. 0 si P= 0,5 a – infinito 1 si P= infinito a 0,5	Relación de una variable binaria dependiente (discreta) y las variables independientes. Probabilidad de que una incidencia se asigne en la clase o etiqueta (variable respuesta).	$P = \frac{1}{(1 + e^{-(\beta 0 + \beta 1X1 + \ldots + \beta pX)})}$

Árboles de clasificación	Relación entre Y y X, de forma que sea posible predecir Y basado en los valores de X. Entrega la etiqueta.	Partición recursiva a través de medidas de ganancia de información o Gini.	Método estadístico no paramétrico. El CART selecciona el corte que conduce al mayor decrecimiento de la impureza (entropía).
Random forest	Los datos se dividen en tantos árboles como se quieran construir.Entrega la etiqueta.	Cultivo de 2 o más árboles de clasificación. Cada árbol da una clasificación y finalmente el bosque elije la clasificación con más votos.	Método estadístico no paramétrico. El CART selecciona el corte que conduce al mayor decrecimiento de la impureza (entropía).
LDA	Método alternativo a la regresión logística. Devuelve la etiqueta y las probabilidades.	Teorema de Bayes, probablidad a priori.	Se modela la distribución de los predictores X para cada clase k por separado: P (Y=k | X=x) Se debe calcular la matriz de desviaciones si hay más de un predictor. Requiere que los predictores estén normalmente distribuidos en cada una de las clases.
SVM	Etiqueta de clasificación.	Hiperplano (plano de decisión que separa conjuntos de objetos) que divide mejor el conjunto de datos en clases. Vectores de soporte, son los puntos de datos más cercanos al hiperplano. Entre ellos se genera un margen; a mayor margen, mejor dicha. La selección del kernel correcto es el resultado.	

Fuente: elaboración propia

4.2 Profundización en el sector del problema de incidentes tecnológicos

Esta etapa se orienta por la observación y la comprensión del sector del problema a través de un primer acercamiento con los datos y entrevistas a analistas y directivos involucrados en el proceso de gestión de incidentes. Además, se detalla el procedimiento actual de clasificación de requerimientos desde la óptica del analista de seguros generales, con lo que también se contemplan

posibles criterios que los llevan a tomar determinadas decisiones de clasificación. Esta etapa arroja un informe técnico de profundización en el sector del problema.

La compañía aseguradora cuenta con un sistema de gestión de incidentes que permite documentar, direccionar, notificar al usuario y resolver los reportados a través de los diferentes canales habilitados para tal fin. Estos incidentes son consignados y documentados en el aplicativo inmediatamente, y son clasificados manualmente por un agente que recibe la llamada; así, son direccionados al equipo gestionador, que se espera que solucione el incidente al indicar la respectiva prioridad. Es importante aclarar que la persona que recibe la llamada se identifica en un nivel 0 de escalamiento, lo que quiere decir que debe tener conocimiento básico para dar solución inicial al incidente. En caso de que esta no cuente con la solución, debe tener un conocimiento mínimo para escalar el incidente al siguiente nivel.

Cuando este se ingresa en el aplicativo, se genera un número de caso, el cual es notificado al usuario que reporta; ello, para que, en caso de ser necesario, pueda hacerle seguimiento. En este momento comienza a contar el tiempo correspondiente a la promesa de servicio que ofrece la mesa de ayuda para dar solución al incidente: cinco días hábiles. Durante este tiempo pueden ocurrir situaciones relacionadas con el escalamiento o la tipificación del incidente, lo que puede llevar a un aumento en los tiempos de respuesta y la insatisfacción del usuario que reporta. Cabe aclarar que esto ocurre cuando el nivel de solución 0 no es efectivo:

- El agente desconoce el nombre del grupo solucionador y debe realizar una consulta a su facilitador.
- El agente asigna el incidente al grupo gestionador incorrecto.

Para comprender con qué criterio clasifica un agente los incidentes, es necesario conocer la distribución de los aplicativos de acuerdo con los dos equipos de solución disponibles, como se muestra en la Tabla 7.

Tabla 7. *Distribución de software por grupo solucionador*

Grupo 0: soportes empresariales
Cliente servidor generales
Cliente servidor hogar
Cliente servidor incendio
Cliente servidor responsabilidad civil
Cliente servidor supyme
Cliente servidor transportes
Cumplimiento web
Módulo integrado de consultas
Perseo
Pyme Express

Grupo 1: soporte automatización
Cotizador - plan empresa
Cotizador empresarial
Suscripción automática generales
Cotizador web sucursal virtual

Fuente: elaboración propia

El nivel de 1 de solución del incidente es el del grupo solucionador de expertos analistas que conocen con detalle la aplicación que presenta el problema y el tipo de error, de quienes se espera que entreguen un diagnóstico del problema reportado con su respectiva solución. En este momento, se identifican otras situaciones que pueden afectar de manera negativa los tiempos de solución y la respuesta al usuario:

- El analista realiza la respetiva revisión de los detalles diligenciados por el agente nivel 0 y determina que el incidente no es de su alcance. En este punto, el analista consume tiempo y procede con la asignación correcta de la etiqueta.

- Cuando el analista determina que el incidente sí es de su alcance, pero necesita validar con un nivel 2 de solución alguna definición o autorización para continuar, se suspende el tiempo de gestión del incidente hasta que le retornen una respuesta. Este punto no pertenece al alcance de este trabajo, dado que el tiempo adicional destinado para la solución no es imputable a una clasificación errónea.

Cuando el analista al que se le asigna el *ticket* correctamente finaliza el análisis de este y lo soluciona o cierra con el respectivo comentario, el usuario es informado automáticamente vía correo electrónico. En el mensaje que se le envía se encuentra la documentación que el analista ingresa para determinar cómo llega a la solución del caso.

Los usuarios que reportan los incidentes relacionados con fallas en el funcionamiento de las aplicaciones de expedición de pólizas presentan continuas quejas al analista de soporte, pues indican largos tiempos de espera de la solución de dichos incidentes, los cuales superan los 30 días. Asimismo, el impacto en la demora para la solución de las fallas a nivel de ventas es significativo, puesto que el cliente final puede declinar la compra de la póliza, ya sea porque no recibe su comprobante o porque al solicitar modificaciones sobre su producto no es posible realizarlas en el tiempo que estas requieren. La venta de pólizas a empresas es un negocio cuyos ingresos varían de acuerdo con el tamaño de la empresa y las coberturas que el cliente desee asegurar; una venta no exitosa puede representar un impacto en la imagen de la compañía y en la disminución de las cifras de ventas según las metas anuales.

Lo anterior se evidencia a través de la información arrojada por la base de datos extraída, donde se muestran los días que un incidente toma en cerrarse (desde que es asignado al equipo correcto y solucionado), en la Tabla 8. Esta presenta la frecuencia de los días que transcurren desde que el incidente se asigna al equipo correcto hasta que es solucionado. Es importante tener en

cuenta que la promesa de servicio son cinco días, y todo lo que supere este tiempo se califica como incumplimiento al cliente.

Tabla 8. *Reporte por días que transcurren desde que se escala el incidente hasta que se cierra o soluciona*

Días cierre	Cantidad	% Frec
De 1 a 5	356	12,5
+ 5	2481	87,5

Fuente: elaboración propia

Se evidencia que, a pesar de que son asignados al equipo correcto, de los incidentes reportados, el 87,5 % demora más de cinco días para dar con una solución. En otras palabras, se incumple la promesa de servicio. Si a esto se suma el tiempo desde que son reportados hasta que se asignan correctamente, el incumplimiento es mucho mayor.

Los equipos gestionadores (grupo 0, de soporte empresarial, y grupo 1, de soporte automatización) están compuestos por analistas con conocimientos especializados y, por lo tanto, son un recurso costoso para la compañía, el cual requiere ser correctamente utilizado. Los miembros de estos equipos son entrevistados para conocer sus principales inconvenientes a la hora de recibir incidentes escalados por el nivel 0 de atención, entre los que se encuentran:

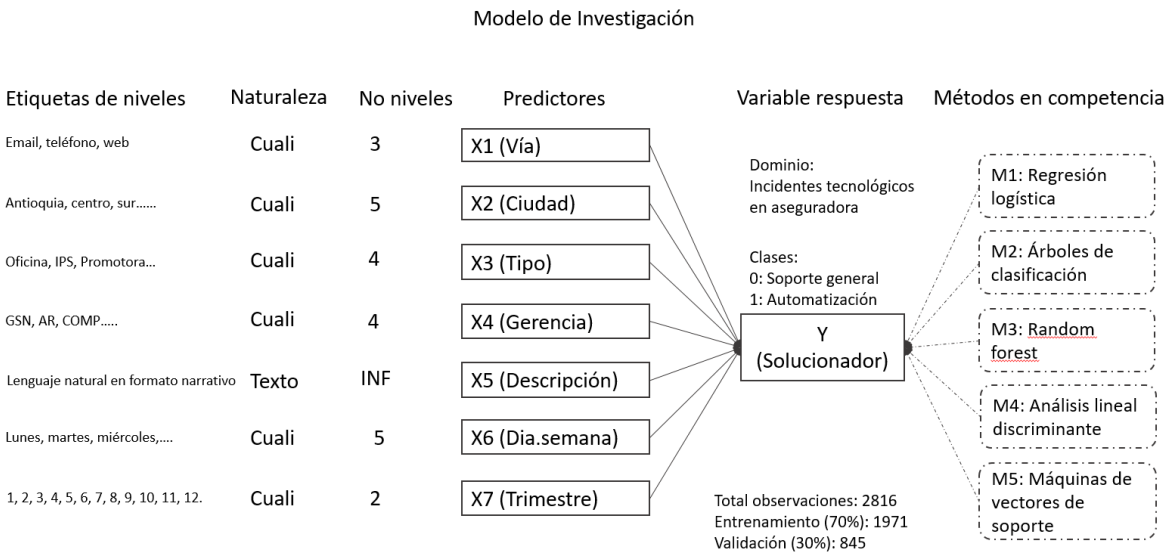
- Debido al volumen y la criticidad de los errores reportados, no es posible revisar la totalidad de incidentes que son escalados a estos equipos de manera diaria; así, si alguno está mal etiquetado, puede demorarse un tiempo adicional mientras alguien del equipo lo toma y lo analiza.

- Cuando el incidente no es etiquetado correctamente, se debe reclasificar. Es posible que los equipos a los que se reasigna el incidente se encuentren en su capacidad máxima y también puede transcurrir un tiempo hasta que puedan tomar el incidente.
- El tiempo de revisión y análisis de un incidente depende de muchas variables; por esto, cuando se etiqueta de manera incorrecta, los equipos que reciben el incidente deben ser hábiles para identificar si la solución está o no a su alcance.
- Es posible que el incidente no esté clasificado correctamente en el equipo, en la subcategoría correspondiente; sin embargo, si pertenece al alcance del equipo, quien lo toma debe asignarle el nombre correcto en el sistema al ingresar a la herramienta de gestión y proceder con la solución.
- En algunos casos, los incidentes no dicen mucho sobre el inconveniente o la falla y los analistas del grupo solucionador deben contactarse con los usuarios para comprender de qué se trata el reporte.

5 Modelo de investigación

En la Figura 2 se presenta el modelo de clasificación de incidentes tecnológicos diseñado para el problema específico planteado en este trabajo, se busca que a través de un enfoque *machine learning* aplicado a una empresa aseguradora, sea posible dar cumplimiento a los objetivos de la presente investigación.

Figura 4. Modelo de investigación para la clasificación de instancias en las clases de la variable solucionador, bajo una comparativa entre cinco métodos de aprendizaje automático



Fuente: elaboración propia

El modelo de investigación diseñado para la clasificación de los incidentes tecnológicos en una compañía aseguradora y su respectivo grupo solucionador, comienza con la selección de los predictores; tal selección incluye variables cualitativas, cuantitativas y en formato texto. Se seleccionan siete predictores, los cuales han de determinar la etiqueta de asignación 0 (soporte general) o 1 (automatización) a través de la aplicación de cinco métodos de aprendizaje supervisado que, según la literatura, son ampliamente utilizados para resolver problemas de clasificación de instancias.

Los datos que alimentan el modelo son 2816 incidentes reportados, donde el 70 % de estos se utiliza como datos de entrenamiento, mientras que el 30 % restante se emplea en el proceso de validación. Finalmente, los métodos a implementar son la regresión logística, los árboles de clasificación, el *random forest*, el LDA y las máquinas de vectores de soporte. Es así como este modelo presenta las variable predictoras y los métodos que serán objeto de la implementación y comparación de resultados posterior.

6 Metodología

6.1 Recolección y limpieza de datos

En esta etapa se define el periodo que contempla las instancias por incluir en la muestra, lo que asegura una cantidad suficiente de instancias correctamente clasificadas en cada categoría o equipo de solución (para este caso son dos categorías 0 y 1).

Del *software* utilizado en la gestión de incidentes de la compañía aseguradora, se extrae un informe que corresponde a los incidentes reportados relacionados con las fallas en los aplicativos. El rango de tiempo para la extracción fue desde el año 2019 (julio, agosto, septiembre, octubre, noviembre y diciembre) al año 2020 (enero, febrero, marzo y abril); este informe corresponde a la población a considerar en el presente estudio. Durante este periodo de tiempo, se asignaron 2816 incidentes a los grupos solucionadores 0 y 1, lo que se configura como la variable respuesta.

Se define un diccionario de variables que se configuran como relevantes de acuerdo con la referencia de la literatura y el aporte para el contexto de la compañía aseguradora; estas han de ser la base para la aplicación de los métodos de aprendizaje supervisado seleccionados. Previo a la división de la muestra, se ejecutan procedimientos de limpieza de datos (imputación, homogenización de etiquetas, etc.), adecuación de tipos de variables (si estas son de factor, numéricas, etc.). También se emplean procedimientos de minería de texto tradicional, basados en frecuencia de palabras o n-gramas, para convertir el texto narrativo en variables independientes cerradas, provenientes de la descripción de los incidentes. Esta etapa arroja un informe técnico sobre el diccionario de variables y el algoritmo de preparación de datos, como se presenta en la Tabla 9.

Tabla 9. *Diccionario de variables*

Nombre original	Nombre nuevo	Descripción	Tipo variable	Valores que toma	Variable respuesta/predictora	Referencia
Created_Via	Via	Se trata del medio a través del cual se reporta el incidente	Categórica	1. Email 2. Teléfono 3. Web	Predictora	Propia
City	Ciudad	Ciudad donde se encuentra ubicado el usuario que reporta la falla	Texto	Apartado, armenia, barranquilla, bogotá, etc	Predictora	Propia
Location	Sucursal	Sucursal autorizada para la venta de pólizas	Texto	Oficinas de asesores, promotoras, etc	N.A	Propia
	Tipo	Tipo de sede donde se vende o se expide la póliza o se realiza la modificación a la misma y donde se detecta la falla en el aplicativo	Texto	Promotora, sucursal, oficina, etc	Nueva predictora	Propia
	Nombre	Nombre comercial de la sucursal.	Texto	Bolivariana, la sabana, Juan botero, etc	Nueva predictora	Propia
Grupo_Gestionador	Gestion	Identificación del grupo que gestiona en la herramienta de gestión de incidentes.	Texto	SAP, Core, etc	N.A	Silva, et al (2018),
	Gerencia	Siglas de la gerencia donde se encuentra el equipo de TI encargado de la gestión. En este caso se realiza la separa	Texto	Seguros, Gsn, sap, etc	Nueva predictora	Propia
	Gestión	Nombre de la coordinación o área que se encarga de dar solución al incidente dentro de la Gerencia indicada	Texto	Centro de servicios, validación de usuario, funcional, autonomía, etc	Nueva predictora	Propia
Sistema	Aplicativo	Nombre del aplicativo afectado	Texto	1. Cliente servidor generales. 2. Cliente servidor hogar. 3. Cliente servidor incendio 4. Cliente servidor responsabilidad civil 5. Cliente servidor supyme 6. Cliente servidor transportes 7. Cumplimiento Web 8. Modulo integrado de consultas 9. PERSEO 10. PymeExpress 11. Cotizador - Plan empresa 12. Cotizador empresarial 13. Suscripción automática generales 14. Cotizador web sucursal virtual	N.A	Silva, et al (2018), Son, et al (2014); Nguyen, et al (2016).
	Tecnología	Se trata de la tecnología utilizada para desarrollar el producto	Texto	Cliente servidor, Cotizador, wrb, etc	Nueva predictora	Propia
	Producto	Nombre como se conoce el tipo de póliza a expedir	Texto	Hogar, Generales, Incendio, etc	Nueva predictora	Propia
Sintoma	Tipo.falla	Descripción general que asigna el agente a la falla reportada	Texto	1. Bloqueos-lentitudes 2. Errores aplicación 3. Manejo aplicación 4. Acceso a aplicación	Predictora	Silva, et al (2018),
Descripcion	Descripcion	Descripción por parte de la persona que reporta la falla y con que se dirige a la mesa de soporte para que lo apoye en la solución de la falla	Texto		Predictora	Paramesh & Shreedhara (2019); Altintas, et al (2014); Sara Silva, et al (2018); Bogojeska, et al (2014); Giurgiu, et al (2017); Al-Hawari y Hala Barham (2019); Gore, et al (2018); Jan, et al (2013); Sulaman (2015); Li y Zhan (2012); Kallis,et al (2019).
Fecha_Asignacion	Fecha.ini	Fecha en que el incidente es escalado al nivel 1 (analistas expertos)	Fecha	Fecha en formato dd/mm/aaaa	Predictora	Silva, et al (2018), Giurgiu, et al (2017)
	Dia.semana	Día de la semana al que corresponde la variable Fecha_asignacion	Texto	Lunes, martes, miercoles, jueves, viernes, sabado, domingo	Nueva predictora	Propia
	Mes	Mes al que corresponde la variable Fecha_asignacion	Texto	Enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre, diciembre	Nueva predictora	Propia
Fecha_Asignacion	Solucionador	Equipo en el que se clasifica el incidente	Texto	0: Equipo empresariales encargados de los aplicativos C/S. 1: Equipo cotizador encargado de los aplicativos web	Respuesta	Al-Hawari y Barham (2019); Altintas, et al (2014); Han y Akbari (2018).
Fecha cierre	Fecha.fin	Fecha en que el incidente es cerrado	Fecha	Fecha en formato dd/mm/aaaa	Predictora	Silva, et al (2018)
Fecha cierre	Dias.cierre	fecha cierre - fecha.ini = Dias.cierre	Texto	# días	Nueva predictora	Propia

Fuente: elaboración propia

6.1.1 Limpieza y tratamiento de datos.

- Se homologa la variable *vía*. Existen dos conceptos: *mail* y *email*, y queda finalmente *email*. Para esta variable no existen datos vacíos.
- Homologar los nombres de las ciudades para la variable ciudad, pues algunos están escritos con tilde, sin tilde, mayúsculas, minúsculas, etc.
- La variable ciudad contiene 10 registros vacíos; se procede con la validación respectiva de la ciudad en la que se reporta el incidente y se diligencian los datos faltantes.
- La variable ciudad se agrupa en zonas del país, así:

Nombre grupo	Ciudades
Antioquia	Medellín, Envigado, Itagüí, Manizales, Apartado, Pereira, Armenia, Rionegro.
Centro	Bogotá, Bucaramanga, Cúcuta, Espinal, Ibagué, Neiva, Santander, Tunja, Villavicencio.
Norte	Barranquilla, Cartagena, Montería, San Gil, Santa Marta, Sincelejo, Valledupar.
Valle del cauca	Cali, Buga, Cartago, Palmira, Popayán, Tuluá.
Sur, Panamá y Rep. Dominicana	Otras.

Fuente: elaboración propia

- Se crean variables nuevas a partir de variables existentes. Se toman las variables *sucursal*, *gestión* y *aplicativo*, dado que sus nombres contienen información que puede descomponerse y generar nuevas variables predictoras que aporten al modelo de clasificación.
- La variable *gestión*, cuyo origen es compuesto, se divide en dos:
    - *Gerencia*: corresponde a la gerencia donde se encuentra el equipo de TI encargado de la gestión. Asimismo, se realiza una agrupación de gerencias:

GSN = GSN

AR= AR

COMP = COMP

Otras = ADM, AUTON, CORE, GGA, GO, GSC, GSO, IBM, INV, MERCADEO, MOVILIDAD, SALUD, SAP, SEGUROS, SEGUROS DE VIDA, SP, SURA.

- *Gestión*: nombre de la coordinación o área que se encarga de dar solución al incidente dentro de la gerencia indicada.

- La variable *aplicativo* de divide en dos:
  - *Tecnología*: se trata de la tecnología utilizada para desarrollar el producto.
  - *Producto*: nombre como se conoce el tipo de póliza a expedir.

Plan empresa = plan empresa, empresarial, sucursal virtual.

Incendio.

Cumplimiento.

Hogar.

Generales = generales, suscripción automática generales.

Pyme Express.

Transporte.

Supyme.

MICtr.

Otras = Perseo + responsabilidad civil + transporte.

- La variable *sucursal* se divide en dos:
    - *Tipo*: tipo de sede donde se vende o se expide la póliza o se realiza la modificación a esta, y donde se detecta la falla en el aplicativo. Ejemplo: promotora, sucursal, taller, etc.

        Sede seguros = autos Sura, Bogotá, seguros Sura.

        Otras sedes propias = IPS, regional ARL, regional EPS, taller, oficina.

        Promotoras = promotoras.

        Sedes no propias = centro comercial, pendiente por asignar, sucursal.
    - *Nombre*: nombre comercial de la sucursal.
- La variable *tipo.falla*:
    - Errores aplicación.
    - No puede ingresar.
    - Otras = manejo aplicación y bloqueos lentitudes.
- La variable *finde*:
    - *Día.semana* = se eliminan registros de sábado y domingo.
- La variable *días.cierre* contiene nueve registros cuyo valor es inconsistente, pues esta es el resultado de la resta entre la fecha que se asigna el incidente al grupo gestionador y la fecha en que se cierra el incidente. Esto se explica porque el sistema permite elegir fechas de cierre sin validar que no sean anteriores a la fecha de asignación. Se procede a igualar la *fecha.ini* y *fecha.cierre* para estos casos.

6.2 Análisis univariado

A continuación, se presenta el análisis de cada una de las variables que se consideran en el modelo de clasificación de incidentes, la herramienta utilizada para realizar dicho análisis es RStudio. Algunas de las variables son de tipo numérico y otras de tipo categórico.

6.2.1 Variables numéricas

- *Fin.cierre*:

  En la Tabla 10, se presenta el resumen estadístico de los datos de la variable.

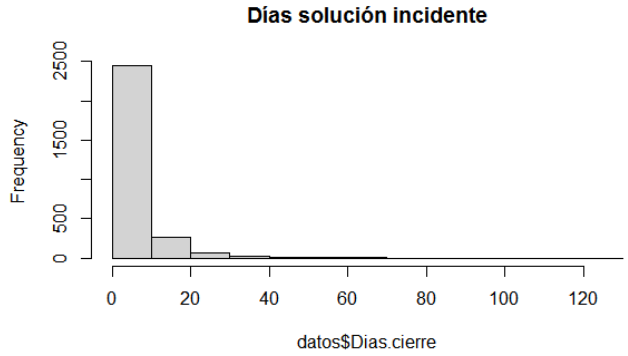
  Tabla 10. *Resumen estadístico*

Min	1st Qu	Median	Mean	3rd Qu	Max.
0,000	2000	4000	6001	7000	123 000

Fuente: elaboración propia

Se observa que la mayor cantidad de observaciones se encuentran dentro del rango 0-7 días. Asimismo, el valor máximo que toma la variable es 123 días; y el mínimo, 0 días. El 75 % de los datos tiene fecha de cierre menor o igual a 7 días. Lo anterior también se puede evidenciar en la Figura 5, que muestra, a través de un histograma de frecuencias, los días que transcurrieron entre la recepción y el cierre o la solución del incidente tecnológico.

Figura 5. Histograma de frecuencias variable *fecha.cierre*



Fuente: elaboración propia

La variable *fin.cierre* indica que los equipos de tecnología reaccionan a los incidentes que son asignados de una manera rápida y de acuerdo con la promesa de servicio, pero quedan algunos que no son cerrados según lo esperado por el cliente. La asignación incorrecta del incidente al equipo solucionador impide el cumplimiento de la promesa de servicio al cliente y las continuas quejas por insatisfacción.

6.2.2 Variables categóricas.

- *Vía*: se trata del medio a través del cual se realiza el reporte del incidente por parte de los usuarios. En la Tabla 11, se presenta la distribución de frecuencias de la vía a través de la cual los usuarios reportan las fallas que presentan los aplicativos.

Tabla 11. *Frecuencia de reporte por tipo de vía*

Vía	Cantidad	% Frec
Email	1820	65
Teléfono	749	27
Web	253	8

Fuente: elaboración propia

En la anterior se observa que la vía de reporte de fallas más utilizada por el usuario es el correo electrónico (65 %). Es importante mencionar que los usuarios, a través de este medio, adicionan archivos tipo imagen para evidenciar los errores generados al realizar la venta de pólizas y los respectivos movimientos de esta. Este hecho es consistente con lo que indica el analista de soporte encargado, puesto que quienes clasifican los incidentes no tienen un contacto directo con los usuarios. En el correo electrónico, el usuario describe la falla presentada.

- Ciudad:

  En la Tabla 12, es posible observar la frecuencia de reporte de incidentes por ciudad.

  Tabla 12. *Frecuencia de reporte por ciudad/zona*

Ciudad/zona	Cantidad	% Frecuencia
Antioquia y Eje Cafetero	1377	49
Centro	869	31
Valle del cauca	269	10
Norte	244	9
Otras	57	1

Fuente: elaboración propia

Se evidencia que las ciudades o zonas donde se encuentran las sucursales de venta más grandes de la compañía, Antioquia y centro (80 %), son las que mayor volumen de incidentes reportan a la mesa de ayuda.

- Tipo:

En la Tabla 13, se presentan los tipos de sedes desde donde se reporta la falla en el aplicativo.

Tabla 13. *Frecuencia tipo de sede origen de reporte*

Tipo de sede	Cantidad	% Frec
Sede seguros	1427	51
Otras sedes propias	510	18
Promotoras	775	28
Otras sedes no propias	104	3

Fuente: elaboración propia

A partir de lo observado en los datos, el reporte de incidentes proviene de las sedes donde se comercializan exclusivamente las pólizas u otros productos propios de la compañía de interés en este trabajo; es decir, sede seguros, otras sedes propias y promotoras (97 %). Lo anterior es

consistente con el mayor uso de los aplicativos y el hecho de que se presente la mayor cantidad de fallas susceptibles de reporte a la mesa de ayuda.

- Gerencia:

La Tabla 14 muestra la frecuencia de reporte de incidentes por gerencia de la compañía aseguradora.

Tabla 14. *Reporte por gerencia*

Gerencia	Cantidad	% Frec
GSN	1772	63
AR	483	17
COMP	157	6
Otros	404	14

Fuente: elaboración propia

Se observa que el 80 % de los registros corresponde a las gerencias AR (administración de riesgos), GSN (gerencias de negocios empresariales) y competitividad empresarial. Desde el mes de febrero de 2020, la mesa de soporte incluye una nueva etiqueta a esta variable; su nombre es competitividad empresarial. Estas gerencias son las encargadas de la contratación y el pago por el soporte técnico de los aplicativos y sus respectivos equipos de solución de fallas a los que se asignan los incidentes.

- *Tipo.falla*:

En la Tabla 15, se observa la frecuencia por tipo de falla, es decir, la categoría que selecciona el agente para definir el tipo de error reportado.

Tabla 15. *Frecuencia tipo de falla reportada*

Tipo de falla	Cantidad	% Frec
Errores aplicación	2651	94
Otras (manejo aplicación y bloqueos lentitudes)	45	2
No puede ingresar	120	4

Fuente: elaboración propia

Se puede determinar que el 94 % de los registros de incidentes tecnológicos corresponde a errores reales de la aplicación y no a temas reportados por desconocimiento o capacitación (manejo de aplicación), calidad de redes o servidores o accesos a dichas aplicaciones. Dado que las aplicaciones interactúan con muchas otras aplicaciones en la compañía, es frecuente encontrar que se desencadenen fallas masivas en diferentes aplicativos; en estos casos, los equipos de tecnología deben escalar el incidente al equipo solucionador adecuado, que puede estar en el alcance de otras gerencias.

- *Día.semana*:

  En la Tabla 16, se presenta la frecuencia de reporte de incidentes por día de la semana.

Tabla 16. *Reporte por día de la semana*

Día.semana	Cantidad	% Frec
Lunes	549	19
Martes	599	21
Miércoles	524	17
Jueves	594	21
Viernes	550	22

Fuente: elaboración propia

Todos los días se observa un flujo constante del reporte a la mesa de ayuda. Se evidencia que los fines de semana los reportes bajan considerablemente, pues son días laborales solo para el canal de venta *retail* y otros que normalmente no presentan grandes volúmenes. Adicionalmente, la mesa de ayuda no cuenta con un equipo de analistas especializados los fines de semana para dar solución a los incidentes reportados durante dichos días.

- Trimestre:

  La Tabla 17 presenta la frecuencia de reporte por mes del año.

  Tabla 17. *Reportes por semestre*

Trimestre	Cantidad	% Frec
AGO-DIC	1699	60
ENE-JUL	1117	40

Fuente: elaboración propia

El trimestre del año en el que se concentra el mayor número de reportes de incidentes es AGO-DIC. Durante estos meses, aumenta el volumen de legalizaciones de ventas para cierre de año por política de la compañía. En este periodo deben cerrarse todos los negocios realizados en el año en curso para que los informes anuales reflejen las cifras de ingresos reales.

- Solucionador:

En la Tabla 18, se presenta el comportamiento de la frecuencia de reporte de la variable *respuesta* por equipo solucionador.

Tabla 18. *Reporte por equipo solucionador*

Solucionador	**Cantidad**	**% Frec**
0	1424	51
1	1392	49

Fuente: elaboración propia

Se puede observar una distribución muy pareja de los incidentes a ser solucionados por cada equipo de tecnología asignado a la tarea de cierre de incidente. Esto se debe a que se cuenta con aplicativos de tecnología antigua (soportes empresariales 0) y aplicativos de tecnología nueva (cotizador 1) para la expedición y modificación de pólizas.

6.3 Entrenamiento de los métodos por comparar usando R

Se programa en R, a través del entorno RStudio, funciones específicas para el entrenamiento de modelos de clasificación de requerimientos para la suscripción de pólizas de seguros por nivel de escalamiento. Una vez programadas, se realizan los controles de calidad, con el fin de detectar y eliminar eventuales anomalías. Adicionalmente, el entrenamiento considera el análisis de parámetros de refinamiento (ejemplo: poda en árboles de clasificación) hasta dejar a punto cada método, listos para competir, guiados por la observación del desempeño de cada uno en la muestra de entrenamiento, con el uso de métricas de la matriz de confusión. Esta etapa arroja un informe técnico del entrenamiento de los modelos, incluyendo los parámetros de refinamiento y los algoritmos diseñados.

Para entrenar y probar los métodos seleccionados, se divide el conjunto de datos de entrenamiento en dos subconjuntos. Un subconjunto (70 %) se utiliza para entrenar el clasificador y el subconjunto restante (30 %) se utiliza para la validación del método. Este proceso evita el sobreajuste, debido a que los conjuntos de entrenamiento son independientes del conjunto de prueba. Se utiliza el entorno RStudio con el fin de realizar el entrenamiento de los métodos, específicamente el paquete Caret (classification and regression training, Kuhn, 2016). Caret es un conjunto de funciones que intentan agilizar el proceso para crear modelos predictivos y de clasificación; algunas de las principales funcionalidades de este son las siguientes:

- División de datos.
- Preprocesamiento.
- Selección de características.
- Ajuste del modelo mediante muestreo.
- Estimación de la importancia de las variables.

Entre las ventajas del uso del paquete Caret se encuentran:

- Utilización de código unificado para aplicar reglas de clasificación muy distintas, implementadas en diferentes paquetes.
- Contiene funciones específicas para dividir la muestra en datos de entrenamiento y datos de test o para ajustar parámetros mediante CV.
- Proporcionar una interfaz uniforme a las funciones en sí, así como una forma de estandarizar las tareas comunes (como el ajuste de parámetros y la importancia de las variables).
- Proporciona la métrica de precisión del método utilizado y no el error o la proporción de clasificaciones incorrectas.

En la Tabla 19, se presenta un informe de los parámetros utilizados en los métodos a comparar en el presente estudio:

Tabla 19. *Parámetros de los métodos a comparar*

Método	Valor del método	Parámetros de ajuste paquete Caret
Regresión logística	Glm	Ninguno.
Árboles de clasificación	rpart (modelo CART)	Cp (parámetro de complejidad). R-cuadrado total debe aumentar en cp en cada paso. La función principal de este parámetro es ahorrar tiempo de cálculo al eliminar divisiones que obviamente no valen la pena. Se utiliza un Cp = 0,01.
Random forest	Rf	Los siguientes parámetros vienen predeterminados en rpart. minsplit = 20, minbucket = 7, maxcompete = 4, maxsurrogate = 5, usesurrogate = 2, surrogatestyle = 0, maxdepth = 30, xval = 0. • Mtry = 2 (número de variables muestreadas aleatoriamente como candidatas en cada división). Jprave (2018) sugiere el cálculo como raíz cuadrada del número de columnas de la matriz de datos -1. • Ntree (número de árboles para crecer). Esto no debe establecerse en un número demasiado pequeño para garantizar que cada fila de entrada se predice al menos unas cuantas veces. Se prueba con 5, 10 y 20. Mejor desempeño del modelo con ntree, 50.
LDA	Lda	Ninguno.
SVM	svmLinear2	Cost (costo); se selecciona un costo de 10 y 15. Se observa que arroja la misma precisión.

Fuente: elaboración propia

En la Tabla 19, en la columna “valor del método”, se presenta el nombre del método para la función Caret y los parámetros de ajuste que requiere el método para entrenar los datos.

7 Resultados

Se presentan los resultados a través de la comparativa entre los métodos de *machine learning* objeto de estudio. Dicha comparación se lleva a cabo bajo un enfoque experimental, al usar todas las métricas de la matriz de confusión (precisión, sensibilidad, etc.). Así, para cada variable respuesta, se reportan los resultados comparativos y se estudia también, según las necesidades prácticas de la organización, qué métricas son las más pertinentes para tomar decisiones de elección del mejor método. De igual forma, se considera la eficiencia computacional en busca de un balance relativo entre eficacia y eficiencia (Kotsiantis, 2007). Esta etapa provee un informe técnico sobre la estrategia de comparación y los resultados tabulares y gráficos, así como los algoritmos diseñados. El objetivo es asignar el incidente al equipo solucionador correcto, al utilizar los cinco métodos seleccionados y las variables definidas. Se establece la presentación de los resultados a partir de la definición de tres escenarios:

- Escenario 1: entrenamiento del modelo al utilizar solo las variables estructuradas más importantes seleccionadas y los cinco métodos de aprendizaje supervisado.
- Escenario 2: entrenamiento del modelo al utilizar solo las variables no estructuradas (matriz de palabras extraídas de la variable, descripción a través de la minería de texto) y los cinco métodos de aprendizaje supervisado.
- Escenario 3: entrenamiento del modelo al combinar las variables estructuradas y no estructuradas más importantes.

A continuación, se presentan los resultados para cada escenario y su respectivo análisis de sensibilidad.

7.1 Escenario 1: entrenamiento del modelo al utilizar solo las variables estructuradas más importantes seleccionadas y los cinco métodos de aprendizaje supervisado

En la Tabla 20 es posible identificar las variables significativas para cada modelo.

Tabla 20. *Variables relevantes para cada modelo*

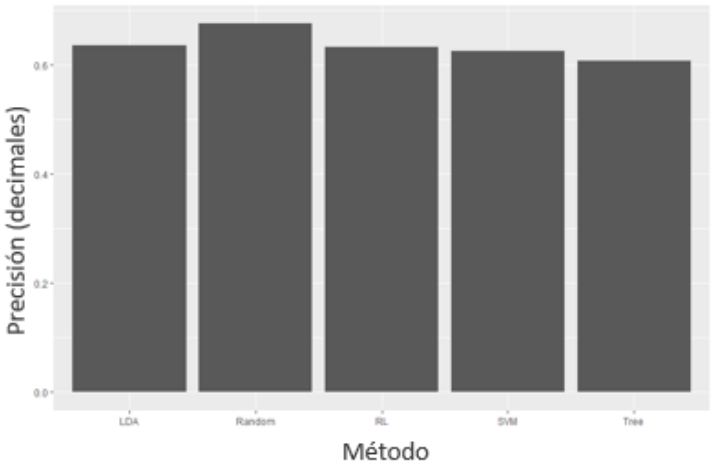
Método	Variables
Regresión logística	Vía (teléfono), tipo (promotora), tipo (sedeseguros), gerencia (GNS), gerencia (otras), ciudad (centro).
Árboles de clasificación	Tipo (promotora), vía (teléfono), tipo (sedeseguros), gerencia (otras), ciudad (centro), tipo (sedes no propias), gerencia (COMP).
Random forest	Tipo (sede seguros), vía (teléfono), tipo promotora, gerencia (otras), ciudad centro, gerencia (COMP), gerencia (GSN), trimestre (EN_JUL).
LDA	Tipo, vía, gerencia, ciudad, día.semana.
SVM	Tipo, vía, gerencia, ciudad, día.semana, trimestre.

Fuente: elaboración propia

En la Tabla 20, se observa que el tipo de sede desde la que se reporta el incidente es relevante para el modelo, así como la vía de reporte, la ciudad y la gerencia a la que pertenece; esto es coherente con la opinión del experto, puesto que se evidencia que las sedes con mayor cantidad de reportes son las promotoras. Es importante mencionar que puede tratarse de fallas no reales que corresponden a temas de capacitación y manejo de la herramienta. De acuerdo con lo indicado por las ventas anuales de la compañía, la ciudad de Bogotá es donde se vende el mayor volumen de pólizas; por tanto, desde esta provienen muchos de los reportes. Así, para este escenario se define el entrenamiento del modelo con la utilización de las variables estructuradas *vía*, *ciudad*, *tipo*, *gerencia*, *día.semana* y *trimestre*, por lo que se evaluará cada uno de los cinco métodos utilizando estas mismas variables.

La *gerencia* es una variable que permite focalizar los aplicativos objetos de reporte, por lo que también es considerada relevante para los modelos estudiados en la clasificación de incidentes. Luego se realiza la selección de las variables estructuradas que más aportan al modelo; para ello, se presenta la Figura 6, donde se observan los resultados de la precisión de los métodos entrenados, donde solo se incluyen las variables de tipo estructurado. Para este caso, en el que los datos tienen una distribución similar en las etiquetas, se utiliza la precisión como estadístico de evaluación.

Figura 6. Resultados de precisión comparativa con variables estructuradas



Fuente: elaboración propia

En la Figura 6, se observa la precisión de cada método con el fin de realizar una comparación de su desempeño, para este escenario el método con mejor desempeño para los datos de entrenamiento fue el *random forest*, con 0,6748; mientras que el de los árboles de clasificación fue el de precisión más baja, con 0,6068. La precisión en todos los métodos se encuentra por encima del 60 %.

En la Tabla 21, se presentan las métricas de sensibilidad (proporción de positivos reales) y especificidad (proporción de negativos reales) de los métodos utilizados para exponer los resultados. La sensibilidad y la especificidad son dos valores que indican la capacidad del modelo para diferenciar los casos positivos de los negativos. La sensibilidad es la fracción de verdaderos

positivos, mientras que la especificidad es la fracción de verdaderos negativos. Para este caso, ambos indicadores determinan cuáles incidentes están correctamente clasificados para cada equipo, así como la capacidad de detectar correctamente los incidentes que corresponden al equipo 0 y los que deben ser asignados al equipo 1.

Tabla 21. *Métricas matriz de confusión comparación métodos*

Método	*Recall*/sensibilidad	Especificidad	Precisión entrena	Precisión valida
Regresión logística	0,724	0,5397	0,6322	0,6627
Árboles de clasificación	0,8878	0,3238	0,6068	0,6438
Random forest	0,8129	0,5356	0,6748	0,6698
LDA	0,7321	0,5367	0,6347	0,6615
SVM	0,817	0,4318	0,6251	0,6651

Fuente: elaboración propia

De la Tabla 21, se muestra la precisión obtenida con cada método tanto para los datos de entrenamiento como validación. Así mismo se presentan otras métricas que aporta la matriz de confusión que permiten observar que los incidentes de la muestra que se clasifican como de los grupos 0 o 1, al ser realmente de estos, son altos para los árboles de clasificación, con un 0,8878; asimismo, pero la especificidad de este mismo método es baja con 0.3238.

7.1.1 Análisis de sensibilidad selección de parámetros *random forest.*

Con el fin de obtener el ntree y el mtry que mejoran el desempeño del método ganador, el *random forest*, se procede a evaluar a través de las variaciones de dichos parámetros en 5, 10, 15, 20, 25, 30, 35, 40, 45, 50. Este procedimiento se realiza sobre la muestra de entrenamiento.

7.1.2 Variación parámetro ntree método Random Forest

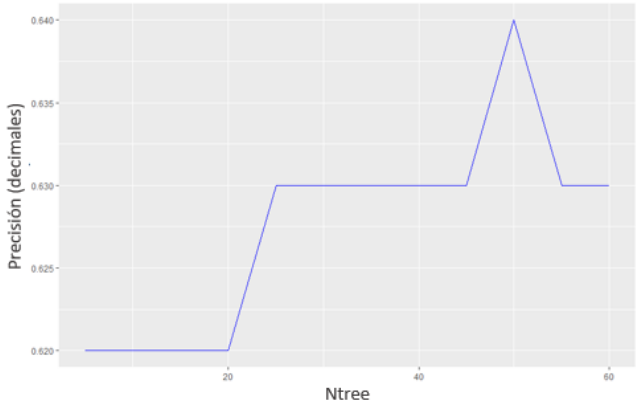
Se presenta la variación de ntree para el método de mejor desempeño, fijando mtry= c(sqrt(ncol(datos2) y ntree tomando los siguientes valores 5,10,15,20,25,30,35,40,45,50,55,60. En la Tabla 22, se presenta el resultado de la precisión del modelo entrenado con *random forest*, bajo estas condiciones. El objetivo es tomar la decisión del parámetro que mejore la eficiencia del modelo.

Tabla 22. *Precisión para variaciones de ntree y grafico asociado*

Ntree	Precisión
5	0,62
10	0,62
15	0,62
20	0,62
25	0,63
30	0,63
35	0,63
40	0,63
45	0,63
50	0,64
55	0,63
60	0,63

Fuente: elaboración propia

Grafico Variación ntree



Fuente: elaboración propia

Se selecciona el ntree = 50, pues es el valor con el que *accuracy* toma su máximo valor de 0,64. De ahí en adelante, las variaciones de ntree arrojan una precisión menor.

7.1.3 Variación parámetro mtry método Random Forest

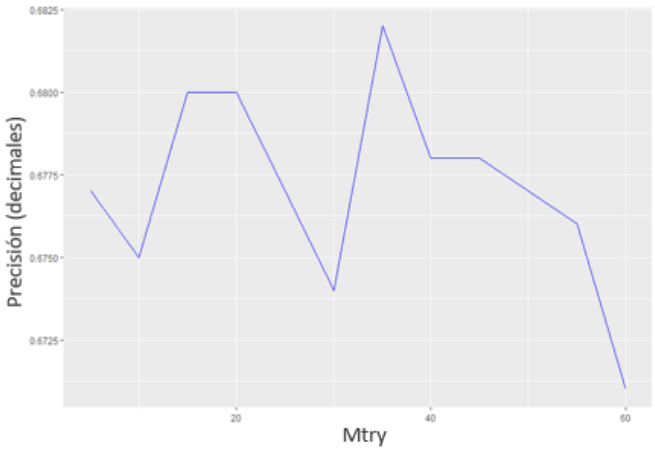
A continuación, de fijaré el parámetro de ntree= 50 definido en el apartado 7.1.2 y se varia Mtry en 5,10,15,20,25,30,35,40,45,50,55,60 con el fin de encontrar el valor en que el método presenta un mejor desempeño.

En la Tabla 23, se presenta el resultado de la precisión del modelo entrenado con *random forest*, el cual varía en mtry y se fija en ntree = 50. El objetivo es tomar la decisión del parámetro que mejore la eficiencia del modelo.

Tabla 23. *Precisión para variaciones de mtry y gráfico asociado*

Mtry	Precisión
5	0,6768
10	0,6753
15	0,6804
20	0,6804
25	0,6773
30	0,6738
35	0,6819
40	0,6778
45	0,6783
50	0,6773
55	0,6763
60	0,6712

Fuente: elaboración propia



Fuente: elaboración propia

Se selecciona un mtry = 35, puesto que es el valor con el que *accuracy* toma su máximo valor de 0,68. De ahí en adelante, las variaciones de mtry arrojan una precisión menor.

7.2 Escenario 2: entrenamiento del modelo al utilizar solo las variables no estructuradas (matriz de palabras extraídas de la variable *descripción* a través de la minería de texto) y los cinco métodos de aprendizaje supervisado

Con el fin de extraer un conjunto de características informativas de la descripción del incidente (variable no estructurada del modelo propuesto), se procede con una limpieza del mismo a través de la minería de texto (Vijayarani et al., 2015). En este modelo, el texto se presenta como un conjunto de palabras, así como lo presentaron (Maksai et al., 2014), quienes lograron obtener un nuevo modelo a partir de las palabras relevantes convertidas en las nuevas variables. En este escenario se busca realizar la comparativa entre los métodos seleccionados utilizando sólo la variable no estructurada en cada caso.

7.2.1 Minería de texto.

Sobre los datos, se procede con una preparación del cuerpo del texto de manera que sea posible establecer la frecuencia de los términos utilizados. Finalmente, se sigue con la creación de la matriz de términos, que se convierte en la variable de los diferentes modelos a probar. Para ello se aplican los siguientes pasos:

a. Remoción de signos de puntuación.

b. Remoción de caracteres especiales.

c. Conversión de texto a minúsculas.

d. Remoción de números.

e. Remoción de *stopwords*.

f. Remoción de espacios en blanco.

g. Conversión de términos a su raíz.

h. Construcción de matriz documentos-términos.

i. Cálculo de frecuencia de términos.

Finalmente, se obtiene el conjunto de palabras, el cual es convertido en un vector de características, como lo sugirieron autores como Muhammad (2015); Altintas y Cuneyd (2014); Silva et al. (2018); Son et al. (2014); (Al-Hawari y Hala (2019); y Kallis et al. (2019). Así, se procede con el entrenamiento, al utilizar los diferentes métodos. Del texto se obtienen 151 términos, con los cuales se corre el modelo inicialmente. En la Tabla 24, se presentan los resultados de los métodos aplicados a las variables no estructuradas, al utilizar las medidas de precisión, sensibilidad y especificidad.

Tabla 24. *Medidas de precisión, sensibilidad y especificidad modelo de variables no estructuradas*

Método	Sensibilidad	Especificidad	Precisión
Regresión logística	0,911	0,8248	0,8681 CI: (0,8523, 0,8827)
Árboles de clasificación	0,8959	0,8065	0,8513 CI: (0,8349, 0,8668)
Random forest	1	0,9969	0,9985 CI: (0,9956, 0,9997)
LDA	0,9221	0,8086	0,8656 CI: (0,8497, 0,8803)
SVM	0,9302	0,8106	0,8706 CI: (0,855, 0,8851)

Fuente: elaboración propia

Considerando los resultados, es posible concluir que la variable no estructurada *descripción* es relevante para obtener la eficiencia en la clasificación de todos los métodos aplicados al modelo, como lo aseguraron (Silva et al., 2018). Con una precisión de 0,9985, el método con mejor desempeño fue el *random forest*, cuya eficiencia fue de 99 %. A una conclusión similar llegaron (Son et al., 2014), quienes indicaron haber obtenido mayor precisión al incluir variables como *asunto* y *cuerpo* del incidente reportado. En el 55 % de los estudios revisados, esta variable fue relevante para lograr un alto desempeño de los métodos. En segundo lugar, de precisión, se encuentra el método SVM, con 87 % para el modelo; este solo incluye la variable no estructurada. Para los demás métodos, se encuentran resultados similares o muy próximos.

El modelo entrenado con la variable no estructurada, y con la utilización del método *random forest*, se evaluó con las siguientes variaciones de cantidad de palabras relevantes (nuevas variables del modelo): 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50. Con esto, se obtuvieron los siguientes resultados:

Tabla 25. *Variaciones de número de palabras incluidas en el modelo*

Variación	Precisión	CI
5	0,8828	(0,8704, 0,8945)
10	0,8988	(0,8871, 0,9097)
15	0,9247	(0,9143, 0,9342)
20	0,9311	(0,9211, 0,9402)
25	0,9435	(0,9344, 0,9518)
30	0,9517	(0,9431, 0,9593)
35	0,9886	(0,984, 0,9922)
40	0,9897	(0,9852, 0,9931)
45	0,9911	(0,9869, 0,9942)
50	0,994	(0,9904, 0,9965)

Fuente: elaboración propia

En la Tabla 25, se evidencia que, con las 15 palabras más importantes en la descripción de un incidente, el modelo es capaz de clasificar con una precisión superior al 92 % igualmente superando a los demás métodos probados. Las siguientes son las raíces de dichas palabras y las posibles palabras generadas, de acuerdo con el contexto de la empresa aseguradora y las fallas reportadas a la mesa de ayuda. Las palabras son:

1. Cotiz = cotizador, cotización, cotizar.
2. Tecnolog = tecnología, tecnológico.
3. Asunt = asunto.
4. Colombi = Colombia.
5. Numer = número.

6. Hog = hogar.

7. Envi = envío.

8. Renov = renovación.

9. Mari = María (nombre de personas que contienen esta palabra).

10. Solicit = solicitud.

11. Empres = empresa.

12. Modif = modificación.

13. Salud = saludo.

14. Plan = plan empresa Sura, plan empresa protegida.

15. Segur = seguro.

Finalmente, presentamos el escenario 3 considerado una combinación de los escenarios 1 y 2.

7.3 Escenario 3: entrenamiento del modelo al combinar las variables estructuradas y no estructuradas

En la Tabla 26, se presentan los resultados de la comparación de los métodos aplicados a un modelo combinado de variables estructuradas y no estructuradas.

Tabla 26. *Comparación de métodos a partir de las medidas de precisión, sensibilidad y especificidad para un modelo combinado de variables estructuradas y no estructuradas*

Método	Sensibilidad	Especificidad	Precisión
Regresión logística	0,9025	0,7489	0,8254 CI: (0,8043, 0,845)
Árboles de clasificación	0,8690	0,8283	0,8486 (0,8285, 0,8671)
Random forest	0,9461	0,8283	0,887 CI: (0,8691, 0,9032)
LDA	0,8894	0,7071	0,7978 CI: (0,7756, 0,8187)
SVM	0,9025	0,7417	0,8217 CI: (0,8005, 0,8416)

Fuente: elaboración propia

De acuerdo con la precisión reportada en la Tabla 26, el método con mejor desempeño continúa siendo el *random forest*, con un 88,7 % de precisión. Esto indica que las variables seleccionadas son aquellas que logran que el desempeño del método sea el esperado.

7.3.1 Variación parámetro ntree método Random Forest.

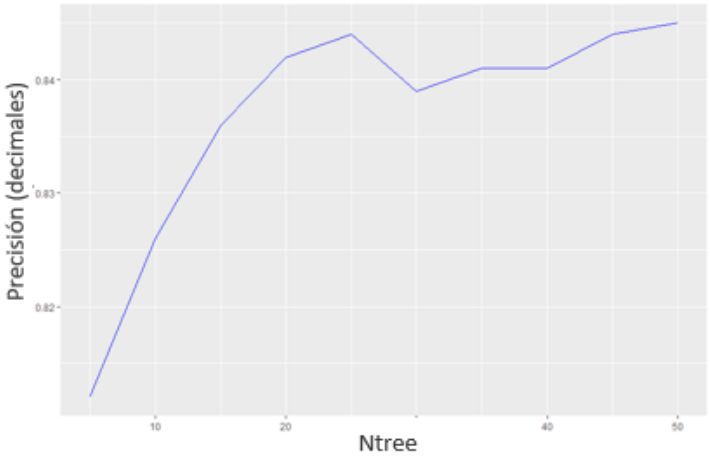
En esta sección se presenta la manera en que se encuentran los parámetros de control del método con mejor desempeño (*random forest*), específicamente se varía parámetro ntree en 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, para la variable no estructurada. Lo anterior permitirá mejorar la eficacia y la eficiencia del modelo. Igualmente, esto busca disminuir el sobreajuste propio del método.

A continuación, se presenta la Tabla 27, que contiene el porcentaje del desempeño del método para cada variación de ntree sugerida. El eje vertical corresponde a los diferentes valores que toma ntree; y el eje horizontal, su respectivo porcentaje de desempeño. Asimismo, se presenta el coeficiente kappa, que toma valores entre 1 y -1: mientras más cercano a 1, mayor el grado de coincidencia entre la clasificación del modelo y la clasificación real de un incidente por el humano.

Tabla 27. *Desempeño de ntree al tomar los valores de 5,10,15,20,25,30,35,40,45,50 y gráfico asociado*

Ntree	Precisión
5	0,812
10	0,826
15	0,836
20	0,842
25	0,844
30	0,839
35	0,841
40	0,841
45	0,844
50	0,845

Fuente: elaboración propia



Fuente: elaboración propia

Se selecciona un ntree = 50, pues es el valor con el que *accuracy* toma su máximo valor de 0,845. De ahí en adelante, las variaciones de ntree arrojan una precisión menor.

7.3.2 Variación parámetro mtry método Random Forest.

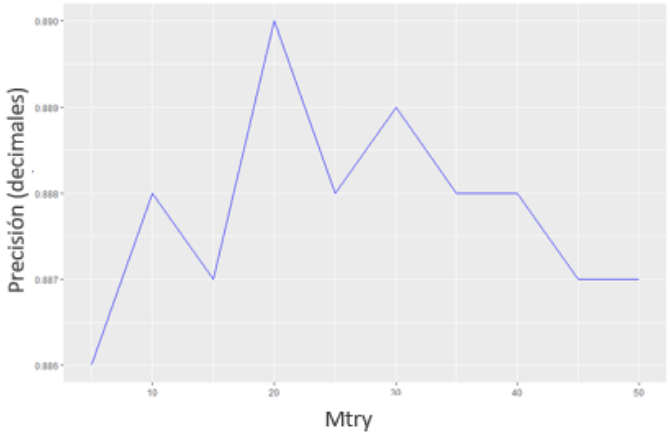
En la Tabla 28, se presenta el resultado de la precisión del modelo entrenado con *random forest*, al variar el mtry en 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 y fijar el ntree = 50. El objetivo es tomar la decisión del parámetro que mejore la eficiencia del modelo.

Tabla 28. *Precisión para variaciones de mtry*

Mtry	Precisión
5	0,886
10	0,888
15	0,887
20	0,890
25	0,888
30	0,889
35	0,888

40	0,888
45	0,887
50	0,887

Fuente: elaboración propia



Fuente: elaboración propia

Se selecciona un mtry = 20, puesto que es el valor con el que *accuracy* toma su máximo valor de 0,890. De ahí en adelante, las variaciones de mtry arrojan una precisión menor.

7.4 Comparación del desempeño de la clasificación del modelo vs. la clasificación del humano (eficacia y eficiencia)

Con el fin de comparar la eficiencia del humano para realizar la tarea de clasificación en contraste con la eficiencia que logra la máquina para realizar la misma tarea, se toma una muestra de 50 incidentes cerrados con su grupo solucionador correcto. Cada incidente es consultado manualmente para saber si fue transferido una o más veces a otros equipos antes de ser solucionado por los equipos de la variable respuesta. Igualmente, se extrae la fecha de creación del incidente y la fecha en que fue cerrado para, finalmente, calcular los días que estuvo abierto desde su creación. Este resultado permite conocer el promedio de días que transcurren para que un agente clasifique

correctamente a la primera vez un incidente reportado a la mesa de ayuda. A fin de calcular la eficiencia del modelo, se plantea la siguiente ecuación:

$$t_{hu} = \rho_1 * \tilde{t}_1 + \rho_2 * \tilde{t}_2 \quad (1)$$

$$t_{hu} = \rho_1 * \tilde{t}_1 + (1 - \rho_1) * \tilde{t}_2 \quad (2)$$

$$t_{hu} = \tilde{t}_2 - \rho_1 \tilde{t}_2 + \rho_1 \tilde{t}_1 \quad (3)$$

$$t_{hu} = \tilde{t}_2 - \rho_1 (\tilde{t}_2 - \tilde{t}_1) \quad (4)$$

Donde:

t_{hu}= tiempo del humano.

ρ_1= probabilidad de clasificar correctamente en cuanto se reporta el incidente.

$\tilde{t}_1$= tiempo medio de clasificar correctamente en cuanto se reporta el incidente.

ρ_2= probabilidad de clasificar incorrectamente en cuanto se reporta el incidente.

$\tilde{t}_2$= tiempo medio de clasificar incorrectamente en cuanto se reporta el incidente.

La eficacia se obtiene a partir del resultado de la precisión que las respectivas matrices de confusión presentan tanto para el humano como para la máquina. Las matrices de confusión se calculan de acuerdo con la definición presentada en el marco teórico.

	Humano	
	0	1
0	31	11
1	6	2
	37	13

Cálculos para el humano:

ρ_1= (31 + 2) / 50 = 33 / 50 = 66 % (precisión)

$\tilde{t}_1$= 6,27 días

ρ_2= 1 - 66 = 34 % (1-precisión)

$\tilde{t}_2$= 11,88 días

Al reemplazar los términos en la ecuación (4), se obtiene:

$$t_{hu} = \tilde{t}_2 - \rho_1(\tilde{t}_2 - \tilde{t}_1)$$

$$t_{hu} = 11{,}88 - 0{,}66\ (11{,}88 - 6{,}27)$$

$$t_{hu} = 8{,}17 \text{ días}$$

En cuanto a los resultados obtenidos por la máquina a través de la función *predict*, se tiene la siguiente matriz de confusión:

	Máquina	
	0	1
0	39	1
1	3	7
	42	8

Cálculos para la máquina:

ρ_1= (39 + 7) / 50 = 46 / 50 = 92 % precisión

$\tilde{t}_1$= 15 minutos

ρ_2= 1 - 92 = 8 (1-precisión)

$\tilde{t}_2$= 15 minutos

Al reemplazar los términos en la ecuación (4), se obtiene:

$$t_{mq} = \tilde{t}_2 - \rho_1(\tilde{t}_2 - \tilde{t}_1)$$

$$t_{mq} = 15\ min$$

A continuación, se presenta un resumen de los resultados obtenidos, tanto para el humano como para la máquina.

Tabla 29. *Eficiencia humano-máquina*

	Eficacia %	**Eficiencia (días/min)**
Humano	66	8,17 / 11 765
Máquina	92	0,01 / 15

Fuente: elaboración propia

De la Tabla 29, es posible concluir que la máquina tuvo un mejor desempeño a nivel de eficacia y eficiencia por encima del humano. La máquina tuvo una precisión de 92 % al clasificar, mientras que el humano tuvo una de 66 %. Adicionalmente, el tiempo que utiliza un humano para realizar la clasificación correcta es, en promedio, de ocho días; en contraste, la máquina no superó los 15 minutos. Al calcular las métricas de sensibilidad y especificidad a partir de la matriz de confusión, tanto para el agente humano como para la máquina, se obtiene lo expresado en la Tabla 30.

Tabla 30. *Sensibilidad y especificidad humano-máquina*

	Sensibilidad	**Especificidad**
Humano	0,74	0,25
Máquina	0,98	0,7

Fuente: elaboración propia

En la Tabla 30, se observan la sensibilidad y la especificidad; estas son las capacidades del modelo o del humano para clasificar incidentes al equipo 0 (sensibilidad) y al equipo 1 (especificidad). Se evidencia que el modelo entrenado presenta una mayor sensibilidad para clasificar incidentes al equipo 0 y al equipo 1, con 0,98 y 0,7; esto, con respecto a los resultados obtenidos para el humano, quien obtuvo 0,74 para clasificar al equipo 0 y 0,25 para clasificar al equipo 1.

8 Discusión

Los métodos probados para entrenar el modelo planteado fueron regresión logística, árboles de clasificación, LDA, *random forest* y SVM. Para todos, y en cada uno de los escenarios planteados, se observó un mejor desempeño del método *random forest*. En la Tabla 31, se observa el resultado de los tres escenarios para dicho método.

Tabla 31. *Comparativa escenarios entrenados con random forest*

Escenario	Sensibilidad	Especificidad	Delta	IC	Precisión
Variables estructuradas	0,8129	0,5356	0,2773	(0,6602,0,6913)	0,6748
Variables no estructuradas	1	0,9969	0,0092	(0,9956, 0,9997)	0,9985
Combinado	0,9461	0,8283	0,1178	(0,8691, 0,9032)	0,887

Fuente: elaboración propia

El escenario que presentó el mejor resultado de precisión fue en el que solo se tuvieron en cuenta las variables no estructuradas, que resultaron de la matriz de términos al aplicar la minería de texto sobre la variable *descripción* de los incidentes reportados. Este escenario obtuvo una precisión del 99,85 %. Ello es consistente con lo que afirmaron Maksai et al. (2014) con respecto a que el desarrollo de un sistema que permite predecir fallas en los servidores a partir de incidentes reportados, donde se utilizan los textos descriptivos en lenguaje natural para clasificarlos, admite tasas de precisión altas en comparación con otros escenarios. Asimismo, Paramesh y Shreedhara (2019) reafirmaron que la narrativa con que se describe el incidente arroja mejores resultados que los encontrados al incluir variables estructuradas en el modelo entrenado. El resultado de precisión al utilizar solo variables no estructuradas es consistente con lo descrito por autores como Altintas y Cuneyd (2014); y Silva et al. (2018). Estos aseguraron que la descripción del incidente juega un papel importante en la clasificación a partir de textos, pues obtuvieron hasta un 86 % de precisión en sus aplicaciones.

En las aplicaciones de los estudios revisados, autores como Giurgiu et al. (2017); y Gore et al. (2018) utilizaron la ingeniería de características para el procesamiento de los textos hasta convertirlos en vectores de palabras. Muhammad (2015) exaltó la importancia de una adecuada preparación de las variables no estructuradas para convertirlas en estructuradas, así como una

adecuada selección de las características de entrenamiento. Para lograrlo, se aplica la normalización del texto con la eliminación de espacios, caracteres especiales, etc. Igualmente, Al-Hawari y Hala (2019) sugirieron la eliminación de elementos no relevantes para la vectorización de las características del texto.

Otros autores, como Kallis et al. (2019), utilizaron herramientas disponibles en la web, como la llamada fastText de Facebook; esto, con el fin modelar el contexto de cada palabra y disminuir la dimensión de las matrices de palabras resultantes. En el trabajo de Silva e al. (2018), se utilizó el criterio de frecuencia de uso de palabras para la selección de aquellas que se incluirían en el vector de características. Finalmente, en el presente trabajo, se seleccionaron aquellas palabras que resultan importantes para el modelo, de acuerdo con el método aplicado y con el uso de la función RStudio Varimport.

Ahora bien, al obtener respuestas a la primera pregunta de investigación, surge la necesidad de profundizar aún más sobre la eficiencia y la eficacia del método probado en el sector estudiado de los seguros. La segunda pregunta de investigación exige la validación del resultado obtenido por el método *random forest* a través de la clasificación de nuevas instancias de forma más precisa y eficiente, y al comparar la actividad del agente humano y de la máquina con su entrenamiento previo. Esta comparativa se logra al realizar la predicción de las nuevas instancias para obtener el resultado de la máquina; y, para el agente humano, se valida la etiqueta de cierre o la solución del incidente para identificar la selección correcta del equipo y los días que estuvo abierto el incidente hasta clasificarse correctamente. Los resultados de la eficacia (precisión) y la eficiencia (tiempo en días que transcurre desde que se reporta hasta que se soluciona el incidente) se presentan en la Tabla 29.

Con respecto a la eficiencia, se toma una muestra de incidentes cerrados y asignados a los dos equipos, y se calculan los días que permanecieron abiertos sin trasferencias (asignados correctamente a la primera vez) y abiertos con transferencias a otros equipos solucionadores (asignados correctamente luego de haber sido reasignados n veces). De manera similar, Al-Hawari y Hala (2019) utilizaron indicadores con el tiempo promedio de asignación de un incidente y el tiempo promedio de su solución para calcular una métrica de eficiencia que permitiera validar si el humano o la máquina ejecutaba de manera eficiente la tarea de clasificación.

Al detenerse en los resultados que arrojan las matrices de confusión, es posible observar la sensibilidad y la especificidad; para este caso, se trata de la capacidad del modelo para clasificar los incidentes al equipo 0 (sensiblidad) y al equipo 1 (especificidad).

En la Tabla 30, se observa que el modelo entrenado presenta una mayor sensibilidad para clasificar incidentes al equipo 0 y al equipo 1, con 0,98 y 0,7. Esto indica que, con el entrenamiento del modelo, la máquina sí adquirió la capacidad de realizar dicha tarea con mayor precisión que el humano.

9 Conclusiones

La utilización del método *random forest* en la clasificación de incidentes tecnológicos para una compañía aseguradora y el uso de la variable no estructurada que presenta la descripción del incidente en lenguaje natural lograron una combinación eficiente para el entrenamiento del modelo presentado. Como lo indican los trabajos revisados, el uso de la variable *descripción* de un incidente conduce a mejores resultados (Silva et al., 2018). A pesar de que la clasificación con este método tomó comparativamente más tiempo que los demás implementados, su rendimiento fue superior; ello, incluso al tratarse de un método que presenta sobreajuste entre la muestra de

entrenamiento y la de validación. Sin embargo, en ninguno de los casos estuvo por debajo del rendimiento de los demás métodos en competencia.

El modelo entrenado a través del *random forest*, al utilizar la variable no estructurada, logró el aprendizaje requerido para presentar mejoras con respecto al desempeño actual del humano en la tarea de clasificación de incidentes; esto puede deberse a la falta de conocimiento, la rotación continua de personal, los análisis errados, etc. Lo cierto es que una clasificación errónea de los incidentes se deriva en retrasos en su solución, lo que afecta al cliente y, evidentemente, las ventas o modificaciones de las pólizas vigentes. Al integrar este modelo en la mesa de servicios de TI, se espera mejorar su capacidad de respuesta y su planeación del recurso para dar paso a la solución oportuna de incidentes, puesto que los altos volúmenes de reporte hacen que esta tarea requiera una mayor velocidad y exactitud.

Es de resaltar la importancia de la limpieza y el tratamiento de las variables no estructuradas a través de la ingeniería de características, pues fue un factor determinante para extraer solo aquellos términos que aportaran a la precisión del modelo. Esta tarea se realiza al normalizar el texto, como lo sugirieron (Gore et al., 2018), al estudiar su semántica y al realizar un procesamiento automático cuando se trata de lenguaje natural.

Es posible concluir que un modelo de clasificación de incidentes para una compañía aseguradora entrenado por el método *random forest* logra ejecutar de una manera más eficiente y eficaz la tarea de clasificación a los equipos solucionadores que un agente humano. Una clasificación correcta y en el menor tiempo posible permite la mejora de los resultados de la operación de la mesa de ayuda, de forma que se pueda realizar una mejor planificación de los recursos utilizados para el proceso de soporte; esto, debido a que, actualmente, el personal especializado contratado debe dedicar un tiempo importante en el análisis y la reclasificación del

incidente en la herramienta. Asimismo, los resultados para la compañía son evidentes, considerando que una respuesta oportuna de las fallas reportadas disminuye la pérdida de negocios con clientes y mejora la imagen.

Finalmente, se confirma la conclusión de Kotsiantis (2007) en cuanto a que la selección de un método de aprendizaje u otro depende de múltiples factores, entre los que se encuentran la exactitud, la velocidad de aprendizaje y de clasificación, los datos de entrada o variables seleccionadas, entre otros. Las múltiples aplicaciones de *machine learning* y los métodos supervisados en el sector de los seguros se consideran un gran apoyo en la toma de decisiones, dado que permiten explorar muchas áreas de las compañías. Entre estas áreas, se encuentran: la gestión de la adquisición de clientes nuevos, la retención de clientes existentes, el análisis de patrones de compra, la asignación de clientes a segmentos de mercado, el predecir compras futuras, el análisis de reclamaciones, el análisis de fraude, la definición de tarifas y la gestión de riesgos propios de la empresa (Umamaheswari y Janakiraman, 2014).

10 Trabajos futuros

Dado que los incidentes son reportados en su gran mayoría por vía correo electrónico, es posible evidenciar que, en muchas ocasiones, los usuarios envían archivos adjuntos; estos pueden ser pantallazos de errores, con lo que se pensaría que, al procesar dichas imágenes, podrían obtenerse nuevos términos que aportaran a una mayor eficacia del modelo. De igual forma sucede para el caso del procesamiento de voz, al momento de realizar el reporte vía llamada telefónica. Con el fin de mejorar la velocidad de entrenamiento del método *random forest* para grandes volúmenes de incidentes, podría aplicarse, como lo sugirió (Kotsiantis, 2007), una partición de la muestra. Aunque el método arroja resultados superiores de desempeño, podría intentarse una combinación con un método de última generación.

Para ello, se sugiere incluir una variable estructurada adicional en el informe extraído y en el modelo de clasificación; dicha variable debe indicar si el incidente tiene más de una transferencia a otros equipos o no, pues la idea es lograr mejores resultados de eficacia cada vez que se entrena el modelo. Actualmente, la información de los incidentes no es analizada en su totalidad, y solo es utilizada para conocer la cantidad de incidentes y el aplicativo que presenta la falla.

Con respecto a los métodos de aprendizaje supervisado entrenados en el presente trabajo, sugerimos en futuros estudios, ampliar las características técnicas y algorítmicas que hacen que el método Random Forest tengo este desempeño en este tipo de contextos y aplicaciones

11 Referencias bibliográficas

Agarwal, S., Sindhgatta, R., & Sengupta, B. (2012). *SmartDispatch:enabling efficient ticket dispatch in an IT service enviroment.* Obtenido de https://www.researchgate.net/publication/254464231_SmartDispatch_Enabling_efficient_ticket_dispatch_in_an_IT_service_environment

Akbar, M., & Jianglei, H. (2018). Vertical Domain Text Classification: Towards Understanding IT Tickets Using Deep Neural Networks. *The Thirty-Second AAAI Conference on Artificial Intelligence*, 8202-8203.

Al-Hawari, F., & Hala, B. (2019). A machine learning based help desk system for IT service management. *Journal of King Saud University – Computer and Information Sciences*, DOI: https://doi.org/10.1016/j.jksuci.2019.04.001.

Altintas, M., & Cuneyd, A. (2014). *Machine learning based ticket classification in issue tracking systems.* Obtenido de https://docplayer.net/1775410-Machine-learning-based-ticket-classification-in-issue-tracking-systems.html

Berk, R. (2017). *Statistical learning from a regression perspective.* Philadelphia USA: Springer texts in statistical.

Berk, R. (2017). *Statistical learning from a regression perspective.* Springer Texts in Statistics.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 1063-1095.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 123–140.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees.* Taylor & Francis.

Bühlmann, P., & YU, B. (2002). Analyzing bagging. *The Annals of Statistics*, 927-961.

Buja, A., & Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 323-351.

de la Fuente, S. (2011). *Análisis Discriminante.* Obtenido de https://www.estadistica.net/Master-Econometria/Analisis_Discriminante.pdf

Giurgiu, I., Wiesmann, D., Bogojeska, J., Lanyi, D., Stark, G., Wallace, R., . . . Hidalgo, A. (2017). On the adoption and impact of predictive analytics for server incident reduction. *IBM J. RES. & DEV*, 23.

Gore, R., Daillo, S., Padilla, J., & Ezell, B. (2018). Assessing cyber-incidents using machine learning. *J. Information and Computer Security*, 341-360.

Gupta, R et al. (2008). *Automating ITSM incident management process.* ICAC.

Gutierrez, J. (2007). *Clasificacion de Imágenes Usando Máquinas de soporte vectorial.* Universidad nacional.

Harrell, F. (2015). *Regression modeling strategies.* Springer.

Hoang, N., Chen, C., & Fang, C. (2016). Automatic classification of traffic incident's severity using machine learning approaches. *Journal The institution of engineering and technology*, 1.

Jan, E., Ayachitula, N., Ni, J., & Zhang, Z. (2013). *A statistical machine learning approach for ticket mining in IT service delivery.* IEEE.

Kallis, R., Di Sorbo, A., Canfora, G., & Panichella, S. (2019). Ticket Tagger Machine Learning Driven. *International Conference on Software Maintenance and Evolution (ICSME)* (págs. 1-4). Computer society.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica, 31*, 249-268.

Kurian, D., Ma, Y., Lefsrud, L., & Sattari, F. (2020). Seeing the forest and the trees: Using machine learning to categorize and analyze incident reports for Alberta oil sands operators. *Journal of Loss Prevention in the Process Industries*, 1-9.

Li, H., & Zhan, Z. (2012). Machine Learning Methodology for Enhancing Automated Process in IT incident Management. *11th International Symposium on Network Computing and Applications* (págs. 1-4). Beijing: University of Posts and Telecommunications.

Maksai, A., Bogojeska, J., & Wie, D. (2014). *Hierarchical Incident Ticket Classification with Minimal Supervision.* IEEE Computer society.

Marcu, P et al. (2009). Towards an optimized model of incident ticket correlation. *IEEE*, 569-576.

Microsoft. (2020). *Microsoft documentación*. Obtenido de https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/create-features

Ministerio de Tecnologías de la Información y las Comunicaciones. (2020). *Home*. Obtenido de www.mintic.gov.co

Muhammad, S. (2015). *Identification of IT incidents for improved risk analysis by using machine learning.* IEEE.

Otero, F., Freitas, A., & Johnson, C. (2012). Inducing Decision Trees with an Ant Colony Optimization Algorithm. *School of Computing, University of Kent, UK*, 3615-3626.

Paramesh, S., & Shreedhara, K. (2019). Automated IT Service Desk Systems Using Machine Learning Techniques. *In Data Analytics and Learning*, 331-346.

Platzi. (s.f.). *Máquinas de vectores de soporte (clasificación y regresión)*. Obtenido de https://platzi.com/contributions/maquinas-de-vectores-de-soporte-clasificacion-y-regresion/

Rolph, A et al. (2007). *Análisis multivariante.* Prentice Hall.

Silva, S., Pereira, R., & Ribeiro, R. (2018). Machine Learning in Incident Categorization Automation. *13th Iberian Conference on Information Systems and Technologies (CISTI).*, 1.

Son, G., Hazlewood, V., & Peterson, G. (2014). On Automating XSEDE User Ticket Classification. *National Institute for Computational Sciences University of Tennessee,*, 7.

Sulaman, S., Weyns, K., & Host, M. (2015). Identification of IT Incidents for Improved Risk Analysis by Using Machine Learning. *Identification of IT Incidents for Improved Risk Analysis by Using Machine Learning* (págs. 1-5). Sweden: ResearchGate.

Umamaheswari, K., & Janakiraman, S. (2014). Role of Data mining in Insurance Industry. *COMPUSOFT, An international journal of advanced computer technology*, 1-6.

Vijayarani, S et al. (2015). Preprocessing Techniques for Text Mining-An Overview. *Int. J Comput Sci Commun*, 7-16.

Yosifova, V., Tasheva, A., & Trifonov, R. (2020). *Most commonly used machine learning algorithms for cybersecurity incident reports classification.* Obtenido de International Scientific Conference Computer Science: http://e-university.tu-sofia.bg/e-publ/files/4484_Paper%20CS%202020.pdf