UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

CATEGORIZACIÓN DE TICKETS DE MESA DE AYUDA PARA UNA EMPRESA DE SERVICIOS INFORMÁTICOS A TRAVÉS DE PROCESAMIENTO DE LENGUAJE NATURAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FRANCO ROBERTO MAGNOLFI SANHUEZA

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
CAROLINA SEGOVIA RIQUELME
BLAS DUARTE ALLEUY

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniero Civil Industrial
POR: Franco Roberto Magnolfi Sanhueza
AÑO:2023
PROFESOR GUIA: Pablo Marín Vicuña

CATEGORIZACIÓN DE TICKETS DE MESA DE AYUDA PARA UNA EMPRESA DE SERVICIOS INFORMÁTICOS A TRAVÉS DE PROCESAMIENTO DE LENGUAJE NATURAL

El trabajo descrito en el presente informe fue realizado en una empresa chilena de servicios informáticos a estudios de abogados llamada Tecnolex. Tecnolex hoy en día se encuentra en una situación en la que la información que poseen algunas de sus áreas no está siendo de confianza. Una de estas áreas corresponde al área de Helpdesk en la cual se realizó el trabajo.

Este correspondía a entregar un sistema que recomendara como categorizar los tickets de soporte técnico que fueran generados por el área. Un ticket cuenta con 2 formas de ser clasificados, por su atributo Tipo Requerimiento y su atributo Categoría, los cuales cuentan con 3 y 14 clases respectivamente. Para visualizar esta situación se tomó un subset de 10.000 tickets, en donde se descubrió que cerca del 40% de estos estaban mal clasificados. También se descubrió que existe una tendencia a clasificar los tickets con un cierto valor, en donde la clase Requerimiento corresponde a aproximadamente el 85% de los tickets existentes. Esto supone un problema para la empresa ya que se basan en esta información para la toma de decisiones del área.

Para lograr clasificar los tickets se hizo uso de tres modelos de machine learning: Regresión Logística, Support Vector Machine y un árbol de clasificación. De los resultados obtenidos, se ha podido concluir que, si existe un gran porcentaje de tickets que actualmente se encuentran mal clasificados, que los modelos de clasificación serían una buena herramienta para clasificar los tickets, y que el modelo que presenta mejores resultados corresponde a la Regresión Logística, el cual permite clasificar los datos de forma correcta un 70% para tanto Tipo Requerimiento como la Categoría de los tickets.

También se logró descubrir que un mayor desbalance en la distribución de los tickets produce peores resultados a la hora de querer clasificar los tickets, en el caso del atributo Tipo Requerimiento mejoro las métricas de las clases con menor cantidad de tickets en cerca de un 30%

A mis abuelos.

TABLA DE CONTENIDO

1. ANTECEDENTES GENERALES

El trabajo de Titulo será realizado en la empresa Tecnolex, la cual es una empresa chilena dedicada a prestar servicios en materia de soporte tecnológico, seguridad informática, gestión de documentos, consultoría en gestión, finanzas y recursos humanos. Ofrecen sus servicios principalmente a estudios de abogados tanto en Chile como en Latinoamérica. Está conformada por 2 sociedades, una limitada y la otra de tipo SPA. Posee una oficina en la comuna de Las Condes y el resto de sus funciones se realizan de forma online. Tecnolex hoy en día cuenta con 60 empleados y les ofrece servicios a 45 clientes.

La Misión de la empresa es la siguiente:

"Proveemos un servicio de excelencia a empresas en materia de soporte tecnológico, seguridad informática, gestión de documentos y en la eficiencia de sus procesos.
Nos esforzamos por conocer en profundidad a nuestros clientes, construyendo una relación de largo plazo, enfrentando juntos los nuevos paradigmas de la transformación digital."
Fuente: Minuta Planificación Tecnolex

La Visión de la empresa es:

"Entregar a nuestros clientes una experiencia de servicio de excelencia que, a través de las últimas tecnologías disponibles, entregue movilidad, flexibilidad y ciberseguridad."
Fuente: Minuta Planificación Tecnolex

Tecnolex cumplió 20 años en septiembre del 2022, y dentro de estos 20 años la empresa se tuvo que reinventar dos veces, buscando nuevos servicios y oportunidades que le entregaba el mercado.

Entre los años 2002 y 2007 entregaba un servicio on site, en el cual cada cliente contaba con sus propios equipos de cómputo, siendo estos administrados por la misma Tecnolex. No existía un área de soporte dedicada a esas funciones, por lo que todo el soporte, gestión de los equipos y servidores era manejado de forma integral.

En el año 2008 Tecnolex arrienda un espacio en un datacenter de GTD y comienzan a trasladar los servidores de los clientes a una zona segura, dónde además se habilitó un enlace a Internet corporativo a través de un Firewall corporativo administrado por Tecnolex. Fue también en este año cuando se creó el área de soporte para usuarios (Helpdesk), estando está separada del área de gestión de servidores.

Para el año 2012 Tecnolex decidió ampliar su carta de servicios, al incrementar su cantidad de recursos locales, permitiendo a los clientes centralizar los servidores, comunicaciones, respaldos e inversiones en tecnología en un único proveedor. Gracias a este nuevo modelo de negocio para el año 2016 Tecnolex logró entregar un servicio completo en términos de seguridad informática a sus clientes.

Durante el año 2020 se realizó un cambio en la forma de entregar los servicios, en donde muchos servicios locales pasaron a ser servicios en nube, utilizando principalmente los recursos de Microsoft y Amazon.

La estructura de Tecnolex está cubierta principalmente por profesionales del área de la computación, contando con un amplio abanico de profesionales, pudiendo cubrir necesidades en los equipos locales en relación a hardware y software, redes locales de los clientes, accesos Wifi, domótica en salas de reuniones, enlaces al datacenter, servidores, respaldos, ciberseguridad, continuidad operativa y gestión de tickets.

Actualmente Tecnolex trabaja con diversas empresas de tecnología y telecomunicaciones, varias de ellas con las que ha llegado a un acuerdo de ser Partners, o les ofrecen servicios de proveedores:

Empresa	Descripción
eSentire	Empresa de tipo World Class, es decir, una empresa que se caracteriza por poseer competencias de alto nivel en todos los elementos de su sistema productivo, la cual se especializa en Ciberseguridad, lo cual le permite a Tecnolex ofrecer servicios de monitoreo provisto por expertos y usar herramientas para investigar ataques, detener brechas de seguridad y anticiparse a posibles vulnerabilidades.
iManage	Software líder en el manejo y gestión de documentos, preferido por más del 70% de los estudios de abogados en el mundo.
Mimecast	Empresa inglesa especializada en el email security, la continuidad y protección de datos del correo electrónico frente a sofisticados métodos de ataque, como por ejemplo phising, impersonation, malware, spam, etc.
GTD Teleductos	Proveedor de Telecomunicaciones y Servicios TIC del grupo GTD.
Amazon WS	AWS es un proveedor de cómputo en la nube, es utilizado por Tecnolex para la replicación de servicios que están en la plataforma local.
Canon Chile	Empresa de origen japonés que provee hardware, software y mantenimiento a las impresoras y scanner de los clientes.

	Permite a Tecnolex contar con un servicio especializado en la gestión de los equipos de impresión y scanners.
Microsoft	Los servicios de Microsoft son utilizados para distintos propósitos, tales como: Servicio de correo electrónico, plataforma de cómputo (Azure) y licenciamiento de software.
Apple	Se cuenta con una licencia de desarrollo de aplicaciones iOS y MacOS.
Automox	Se utiliza para la gestión de actualizaciones y administración de los equipos de usuario.
DUO Security	Se utiliza para la incorporación de un doble factor de autentificación.
Cisco	Se utiliza para la gestión de las redes locales de los clientes.
VMWare	Se utiliza para administrar los servidores de forma virtualizada.
Veeam	Se utiliza para realizar los respaldos y replicación de los servidores virtuales.
Palo Alto Networks	Se utiliza como solución de Firewall corporativo.

Tabla 1: Partners Tecnolex

Gracias a las herramientas entregadas por los proveedores previamente indicados, más las propias capacidades de Tecnolex, esto les ha permitido entregar los siguientes servicios a sus clientes:

Servicio	**Descripción**
Gestión Documental	Organización de documentos e emails a través del uso de una base de conocimiento centralizado y ordenada, lo cual permite administrar accesos, versiones y realizar mejores prácticas.
Time y Billing	Herramienta móvil que permite controlar el trabajo diario de los equipos, de esta forma facilitando la gestión, cobro y facturación de forma eficiente.
Email Security	Como se mención previamente en el informe, protección de su correo electrónico y la información que contiene de cualquier tipo de peligro.
Compliance	Tecnolex es una empresa certificada por la norma ISO 27001, lo que le permite prometer altos estándares de calidad en el cumplimiento de sus compromisos.
Helpdesk	Tecnolex ofreces servicios de soporte tecnológico y mesa de ayuda remota.

Operaciones	La empresa se compromete a asegurar el acceso presencial y remoto a su trabajo a través de un servicio estable y de alta disponibilidad, con un uptime superior a 99,44%.
Ingeniería Legal	Ofrece servicios de integración y de construcción de procesos utilizando aplicaciones desarrollados por profesionales.
Proyectos Especiales I+D	Desarrollo de nuevas tecnologías y aplicaciones.

Tabla 2: Servicios principales Tecnolex

Como se puede apreciar en los servicios indicados, uno de los pilares fundamentales de Tecnolex es el manejo de la información confidencial, siendo una de sus metas conseguir la certificación de la norma ISO 27001, la cual logró conseguir durante el año 2018. La ISO 27001 es un estándar para la seguridad de la información, la cual representa el deber de velar por la confidencialidad, la integridad y la disponibilidad de la información. La norma está dividida en dos partes, en donde la primera se compone de 10 puntos principales que una empresa debería cumplir y la segunda parte los controles que se deberían implementar para controlar el cumplimiento de los 10 puntos de la primera parte.

Con respecto a la participación de mercado de Tecnolex, hoy en día les ofrece servicios a estudios de abogados pertenecientes a los grupos más prestigiosos y mejor calificados de Chile, a la vez que ofreciendo servicios a algunos países de Sudamérica.

Dentro de sus clientes en Chile se encuentra:

- Carey
- Barros y Errázuriz
- Alessandri y Compañía
- Guerrero Olivos
- Albagli Zaliasnik abogados
- CMS
- Larraín abogados
- Prieto y abogados
- Gutiérrez, Waugh, Jimeno y Asenjo
- Vergara Galindo Correa y Abogados

Sus clientes internaciones a los cuales les ofrece servicios corresponden a:

- Vouga y Olmedo abogados (Paraguay)
- Miranda y Amado (Perú)
- Guyer y Regulez (Uruguay)

La principal meta de Tecnolex es que con los servicios que ofrece, la participación de los proveedores con los que cuenta y su certificación ISO 27001, transformarse en el principal departamento de Tecnologías de la Información de sus clientes.

Con respecto a las facturaciones de Tecnolex, actualmente es considerada una "Gran Empresa", obteniendo una facturación promedio mensual sobre los 370.000.000 CLP durante el año 2022.
(Fuente
https://www.sii.cl/preguntas_frecuentes/catastro/001_012_6503.htm)

Para lograr la meta indicada previamente y cumplir con los servicios prometidos a sus clientes, Tecnolex presenta el siguiente organigrama, con una breve descripción de las funciones de cada una de las principales áreas de la empresa.

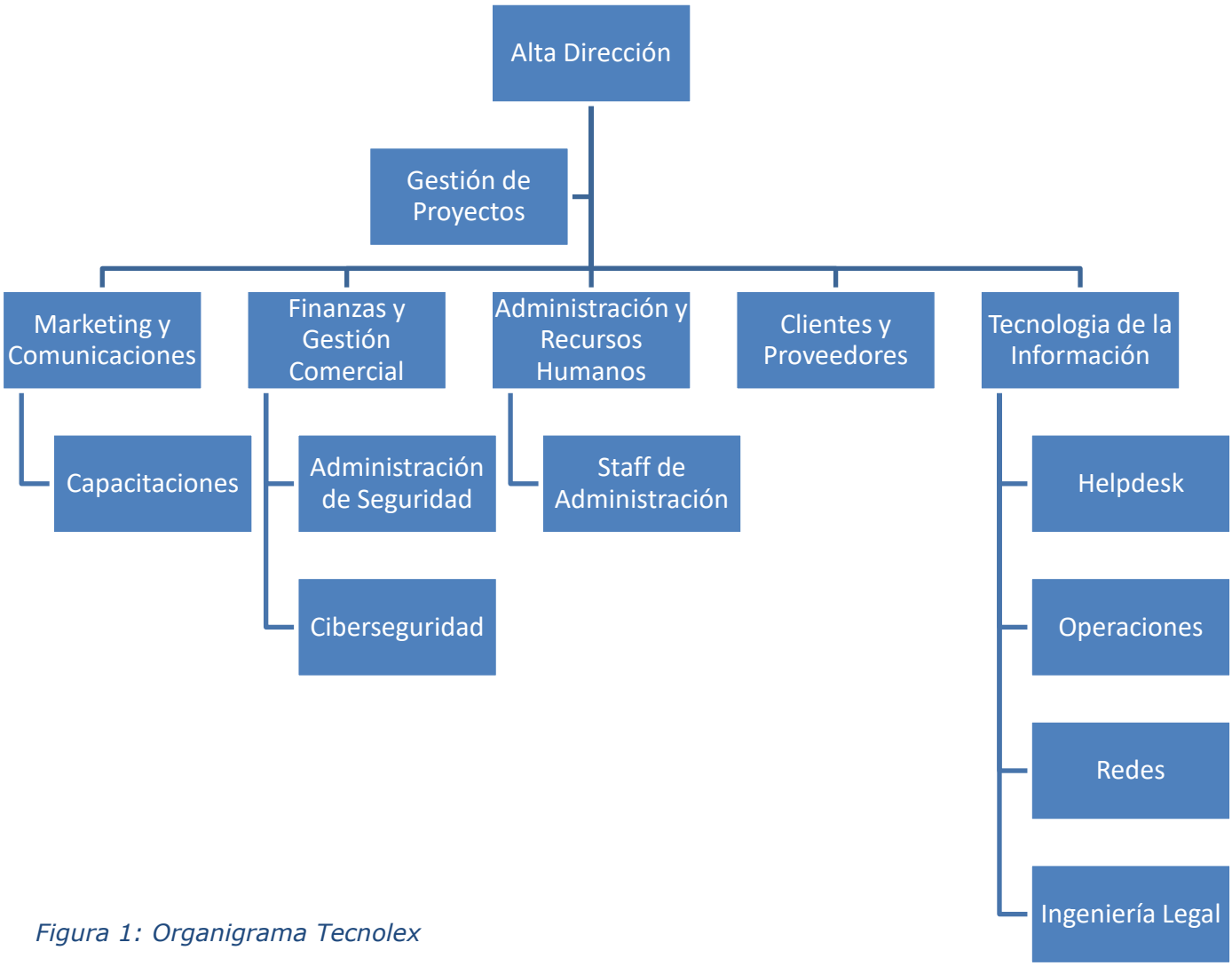

Figura 1: Organigrama Tecnolex

Alta Dirección: Compuesta por los socios y el grupo directivo de la empresa.

Gestión de Proyectos: Corresponde al área de PMO encargada del monitoreo y control de los proyectos realizados por Tecnolex.

Marketing y Comunicaciones: Encargados de la imagen interna y externa de Tecnolex. También son los encargados de coordinar y gestionar las capacitaciones de la empresa.

Finanzas y Gestión Comercial: Administración de los fondos de la empresa y el presupuesto de la misma.

Administración y Recursos Humanos: Gestión de los pagos, gestión administrativa y de los empleados.

Clientes y Proveedores: Gestión de contratos, propuestas y monitoreo de las relaciones con clientes y proveedores de la empresa.

Tecnología de la Información: Gestión de todas las operaciones globales de los procesos de Tecnolex asociados a la gestión de los servicios, manejo de data centers, gestión de las comunicaciones electrónicas, y la administración del área de desarrollo de software.

2. DESCRIPCIÓN DEL PROYECTO Y PLANTEAMIENTO DEL PROBLEMA

Para poder plantear el problema que se desea solucionar, se debe entregar el siguiente contexto del funcionamiento de Tecnolex y de su área de Helpdesk, en donde se realizará el trabajo de título.

Hoy en día, el CEO de la empresa está buscando como una de sus metas aumentar la participación de mercado de Tecnolex, pero sin incrementar el número de empleados con los que cuenta. Una de las solicitudes de la Alta Gerencia a los respectivos jefes de las áreas pertenecientes a Tecnolex es realizar una revisión de la información que Tecnolex posee actualmente, con el fin de verificar si esta es de confianza y realmente representa la situación de la empresa y de sus clientes, a la vez confirmando si se puede hacer uso de esta información para la toma de decisiones a futuro.

Dado lo anterior, cada uno de los jefes de las áreas debe presentar reportes mensuales a la Alta Gerencia, indicando la situación actual de cada una de las áreas. Es en la entrega de estos reportes en donde se presenta el problema que se busca resolver con este trabajo, específicamente para los reportes generados por el área de Helpdesk de Tecnolex.

A continuación, se presenta el funcionamiento y estructura del área de Helpdesk, con el fin de mostrar cual es el problema y porque este existe a la hora de generar los reportes para la Alta Gerencia.

Helpdesk cuenta con la siguiente estructura, la cual se puede visualizar en el organigrama a continuación:

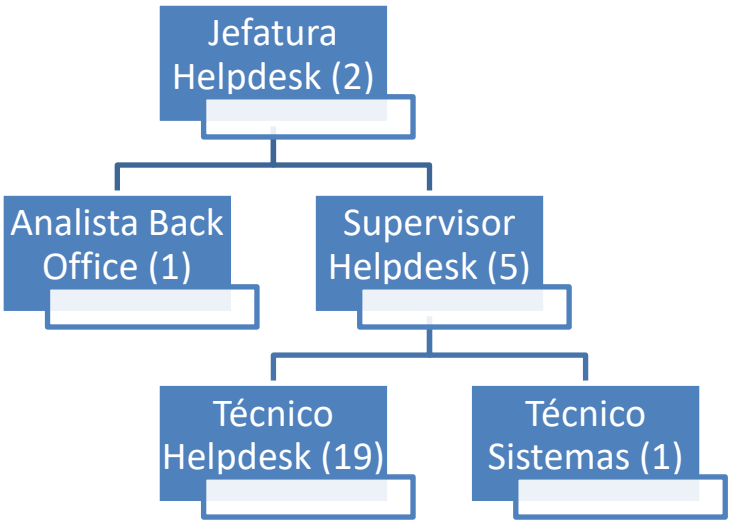

Figura 2: Organigrama del área Helpdesk

Se entrega una breve descripción de cada uno de los cargos del área:

Jefatura Helpdesk: Garantiza el adecuado funcionamiento y calidad de las tecnologías de apoyo a la operación de los usuarios de los clientes para que el desempeño de sus labores sea óptimo, respetar y velar por el correcto uso de los procesos y metodologías definidas por la organización.

Analista Back Office: Gestiona adecuadamente los sistemas y plataformas, controla los equipos que son utilizados por la entidad y brinda los reportes necesarios para posteriores análisis de capacidad por parte de jefaturas.

Supervisor Helpdesk: Supervisa el adecuado funcionamiento y calidad de las tecnologías de apoyo a la operación de los usuarios de los clientes para que el desempeño de sus labores sea óptimo, respeta y vela por el correcto uso de los procesos y metodologías definidos por la organización.

Técnico Sistemas: Provee el adecuado funcionamiento y calidad de los sistemas de los clientes para que el desempeño de sus labores sea óptimo, respeta y vela por el correcto uso de los procesos y metodologías definidos por la organización.

Técnico Helpdesk: Provee el adecuado funcionamiento y calidad de las tecnologías de apoyo a la operación de los usuarios de los clientes para que el desempeño de sus labores sea óptimo, respeta y vela por el correcto uso de los procesos y metodologías definidos por la organización. A los técnicos también se les puede llamar analista.

El área de Helpdesk hace uso de un documento llamado ticket de soporte técnico, que para este informe se nombrará solamente como ticket. Un ticket corresponde a un boleto digital, el cual representa una solicitud de atención por parte de un usuario específico de Tecnolex. Este usuario puede corresponder tanto a un cliente como un funcionario de Tecnolex. Para poder generar y administrar los tickets se hace uso de un sistema de tickets.

Con respecto al sistema de tickets implementado por Tecnolex, el área de soporte actualmente posee un software de origen israelita y su nombre es SysAid.

El sistema de gestión de tickets (SysAid) está integrado por dos sistemas:

- Correo electrónico.
- Call Center.

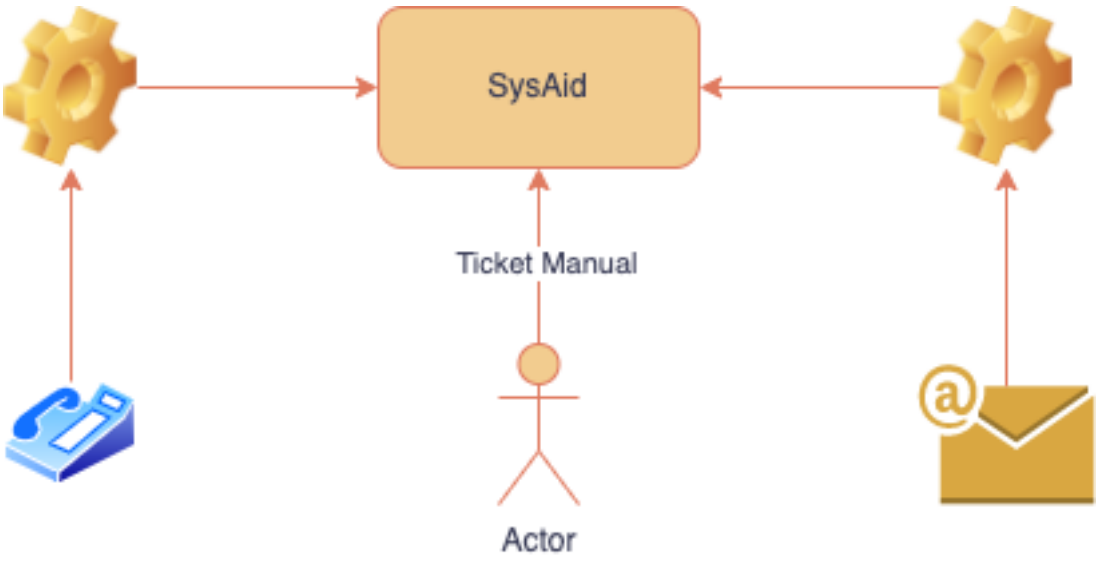

Figura 3: Sistema de gestión de tickets SysAid

Con esto un ticket posee tres orígenes distintos, estos pueden ser email, llamada telefónica o ingreso manual.

Actualmente Tecnolex se basa en un estándar internacional llamado ITIL, este estándar no permite certificar empresas, solo a personas.
El estándar ITIL maneja tres grandes procesos:

- Gestión de incidentes.
- Gestión de problemas.
- Gestión de cambios.

Los tres procesos son gestionados a través de la herramienta SysAid.

Los tickets no solo cuentan con posibles distintos orígenes, estos cuentan con 23 atributos que representan información pertinente al ticket, siendo su origen uno de estos atributos. Algunos de estos atributos son llenados de forma automática por el sistema SysAid, otros se deben llenar de forma manual de parte de los técnicos, y un tercer grupo puede ser llenado de ambas formas, dependiendo de su origen y el estado en el que se encuentre.

En la siguiente tabla se puede apreciar esta lista de atributos:

Atributo	Descripción	Llenado
Reporte	Corresponde al número del ticket	Automático
Estado	Situación la que se encuentra el ticket, el cual puede tomar el valor: • Nuevo • Cerrado • En proceso • Llamada equivoca/ No aplica • Pendiente Proveedor • Pendiente Cliente • Solucionado	Automático inicialmente, pero a medida que cambia el estado se llena de forma manual.
Responsable Tecnolex	Técnico o analista encargado de solucionar la situación descrita en el ticket.	Manual
Grupo	Área de Tecnolex que está a cargo del ticket, esta área puede cambiar a medida que avanza la resolución del ticket.	Automático
Origen	De donde proviene el ticket	Automático
Fecha y hora de apertura	Fecha y hora de apertura del ticket	Automático
Fecha y hora de cierre	Fecha y hora de cierre del ticket	Manual
Tiempo transcurrido	Tiempo transcurrido desde que se abre el ticket hasta hoy	Automático
Tiempo a cierre	Tiempo desde que se abre el ticket hasta que se cierra	Automático
Nombre Solicitante	Nombre de la persona que realiza la solicitud o presente el problema.	Por correo de forma automática. Por llamada telefónica o ingreso manual del ticket de forma manual
Correo Solicitante	Correo de la persona que realiza la solicitud o presenta el problema	Automático
Rol Solicitante	Rol de la persona que realiza la solicitud o presente el problema dentro de su organización	Automático
Estudio	Empresa a la cual pertenece el solicitante. Se asigna el nombre Estudio a este atributo ya que los clientes de Tecnolex son estudios de	Mismo caso que para atributo Nombre Solicitante

	abogados, pero dentro de los valores que puede tomar el atributo se encuentra el mismo Tecnolex.	
Categoría	Categoría del ticket, indicando el tipo de servicio o área de Helpdesk que se puede apreciar en el ticket.	Manual
Sub Categoría	Sub categoría dependiente de la primera categoría del ticket	Manual
Tercera Categoría	Ultimo nivel de categorización del ticket	Manual
Título	Resumen del ticket. Para los correos corresponde al asunto del mismo.	Automática para correos al corresponder a su asunto. Se debe llenar de forma manual para los otros 2 orígenes.
Descripción	Detalle con la solicitud del ticket. Para los correos corresponde al cuerpo del mismo	Mismo caso que para el atributo Titulo
Tipo de Requerimiento	Representa el tipo de situación a la que corresponde el ticket. Puede tomar los valores: • Consulta • Requerimiento • Incidente	Manual
Tutor	Persona encargada del grupo en el cual se encuentra el Responsable Tecnolex.	Automático
Solución	Descripción de la solución aplicada al ticket.	Manual
Ticket padre	Indica todos los tickets extra que se han generado aparte del original.	Automático
Tiempo Respuesta	Tiempo en segundos en que toma atender el ticket y generar primera respuesta al cliente.	Automático

Tabla 3: Atributos Tickets.

A continuación, se presenta el listado de actividades que debe seguir un técnico para el tratamiento de tickets en el sistema SysAid. Se dividirá en los pasos para tratar con tickets generados a través del Call Center y tickets generados por correo.

Los pasos para ingresar al Call Center son:

1. Ingresar a la siguiente URL: https://192.168.30.250/index.php
2. Ingresar a la cuenta con sus datos de Usuario y Contraseña.
3. Existe una parte del portal de Call Center llamada "Agent Number" donde cada técnico debe seleccionar el que le corresponda.
4. En la sección "Extension" del portal debe elegir su cola de llamada, la cual en el área es llamada anexo.
5. Cada técnico es asignado una clave personal la cual debe ser usada para ingresar. Una vez escrita la clave se debe seleccionar el botón "Enter".
6. En el momento en que suene su teléfono y el técnico atienda la llamada, se genera un ticket de soporte. También a partir de este momento el técnico se encuentra conectado al sistema de Call Center.

   Durante una llamada, el técnico puede tomar las siguientes acciones:

   - Take a Break: Se utiliza cuando se deba realizar una pausa del servicio.
   - Transfer: Cuando se requiera transferir la llamada a otro anexo u cola de llamada, con el fin de que la llamada sea atendida por otro miembro del área.
   - HangUP: Esta acción se toma cuando se termina de atender y el técnico debe colgar.
   - End Session: Una vez haya terminado su jornada, el técnico debe hacer click en este botón para cerrar su sesión.

Algunos puntos importantes que debe seguir un técnico de Helpdesk a la hora de responder tickets a través del Call Center:

- Al tomar un ticket se debe cambiar su estado a En proceso.
- Cuando se atiende la llamada el sistema genera un ticket de forma automática, y es trabajo del técnico que está realizando la atención indicar que es el responsable de ese ticket, ingresando su nombre en el atributo "Responsable Tecnolex".
- Si al revisar en el listado de tickets no logra encontrar el que le corresponde, ya sea porque el sistema no lo generó de forma automática

u otro motivo, el técnico debe crear el ticket de forma manual, indicando todos los atributos que le correspondan a ese ticket.

- Una vez se ha realizado la atención, el técnico debe cerrar el ticket indicándolo en su estado.
- Si un ticket queda pendiente por el motivo que sea, se debe agregar actividad en el ticket indicando el motivo informando por correo, y se debe cambiar el estado del ticket a Pendiente proveedor o Pendiente cliente según corresponda el caso.
- Si luego de todas las pruebas realizadas para solucionar un problema atendido por teléfono se requiere asistencia, se debe solicitar esta ayuda a los analistas en terreno a través de correo. Estos correos correspondes a nuevos tickets, por lo que se debe indicar en el atributo "Ticket Padre" cual corresponde al ticket original.
- No debe cortar la llamada en su teléfono, pues esto lo desconecta del Call Center.
- No debe cerrar su sesión hasta que finalice su jornada.

Con el fin de presentar visualmente el proceso por el cual pasan los tickets se presenta la siguiente figura:

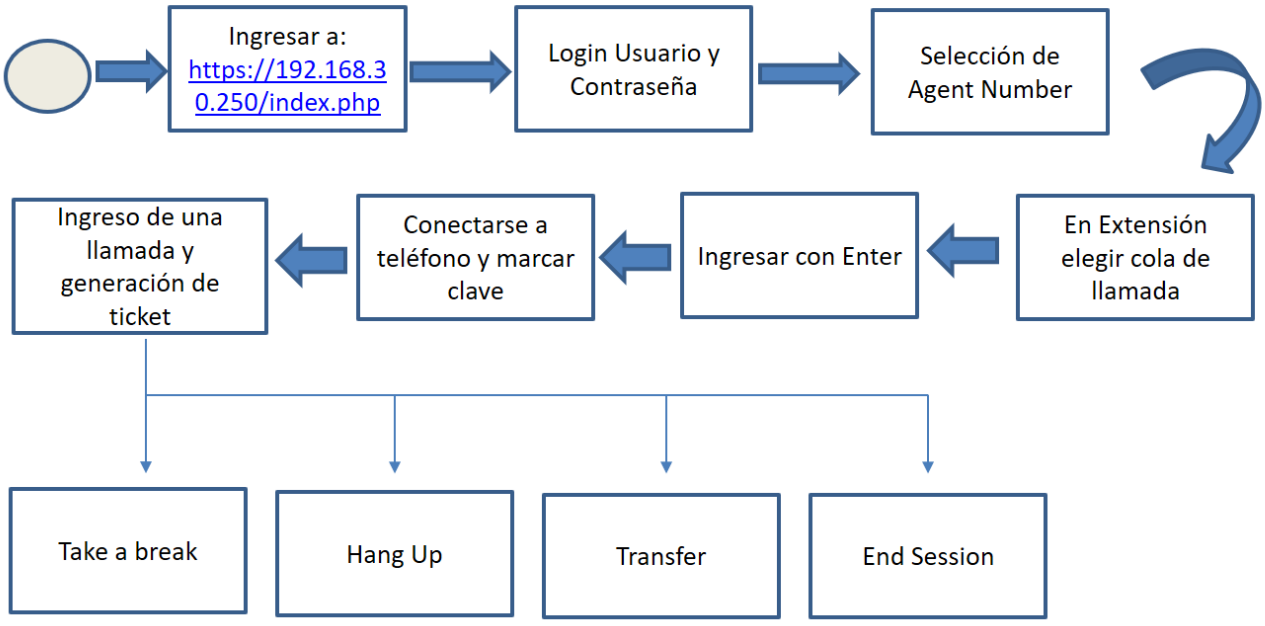

Figura 4: Diagrama creación tickets con origen teléfono

En la figura se pueden apreciar todos los pasos mencionados previamente.

Ahora, para el ciclo de vida de los tickets en el sistema SysAid generados por correo:

1. Se recibe correo con la solicitud o problema del usuario.
2. Se genera ticket en SysAid.
3. Analista o técnico que tome el caso debe asignarse el ticket, indicando en el atributo "Responsable Tecnolex su nombre".
4. Analista que se asignó el ticket debe enviar correo a Helpdesk, agregando en el asunto el número del ticket e informando que tomo el caso. Dentro del correo debe ir un mensaje que diga "Lo veo".
5. Al igual que con los tickets generados por llamadas, en el caso que no se pueda resolver la situación descrita en el ticket, se cambia el estado a Pendiente cliente o Pendiente proveedor cuando corresponda. Se debe enviar otro correo a Helpdesk como respuesta al correo con el mensaje "Lo veo" indicando por que el ticket quedo pendiente. También los técnicos le pueden agregar comentarios a los tickets.
6. Si no se logra solucionar la situación indicada en el correo durante la jornada en la cual fue atendido el ticket, se debe hacer seguimiento al ticket para actualizar tanto vía comentario o correo el estado en el cual encuentra el ticket.
7. Una vez se ha solucionado la situación, se debe agregar la solución al ticket y enviar un correo a Helpdesk danto una descripción de la misma.

Puntos importantes que deben considerar los técnicos:

- Los puntos 1 y 2 son realizados de forma automática por SysAid.
- Los pasos del 3 al 7 son obligatorios y responsabilidad de los técnicos.
- Todos los correos enviados a los usuarios deben ir con copia a Helpdesk.
- La única excepción al caso anterior es cuando la suma de los adjuntos del correo supera los 300KB, caso en el cual solo se le envía el correo al usuario y después se debe reenviar ese correo sin adjuntos a Helpdesk con el mensaje "FYI" (sin comillas).
- Si la persona que envía el correo pregunta por el estado del ticket, esta debe ser respondida a la brevedad posible.
- El ciclo de vida de un ticket que se genere de forma manual debe ser informado igualmente.
- Todos los correos enviados a los clientes deben ser enviados desde Helpdesk (helpdesk@tecnolex.cl).
- Con respecto a las firmas, si se debe enviar un correo interno, se hace uso de una firma corta. En el caso de enviar un correo a un cliente se debe enviar la firma completa.

Para este proceso también se presenta una figura mostrando todos los pasos indicados:

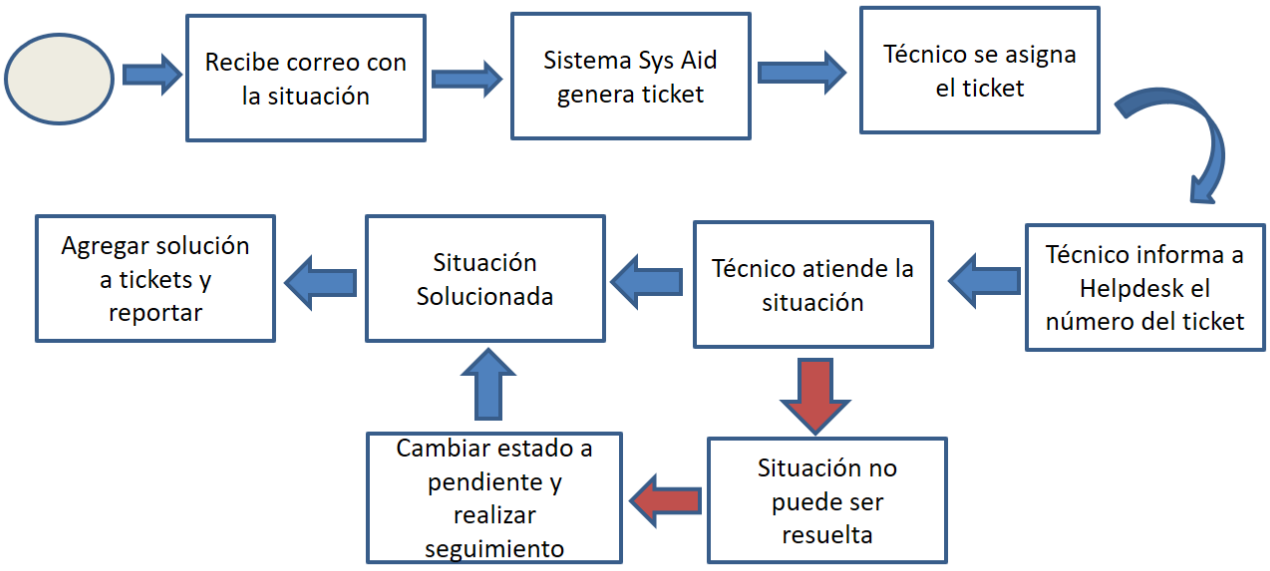

Figura 5: Diagrama creación tickets con origen correo

Ahora, con respecto al problema mencionado al inicio de esta sección, este se encuentra en el llenado manual de los atributos de los tickets, específicamente para los campos Tipo Requerimiento, Categoría, Sub Categoría y Tercera Categoría. Uno de los jefes de Helpdesk, con quien se acordó el trabajo, afirmo que *"Cada vez que reviso un ticket está mal clasificado"*. Con el fin de verificar si esta afirmación es correcta o no, se realizó una revisión manual del campo "Tipo Requerimiento" para una muestra correspondiente a los 10.000 tickets más recientes de Helpdesk, y se encontró que cerca del 40% de los tickets tenía una clasificación asignada incorrecta. Esa diferencia de cerca del 40% se puede apreciar en la siguiente tabla, la cual muestra las distribuciones para las correspondientes clases del atributo "Tipo Requerimiento" antes y después de haber realizado la revisión manual.

Etiqueta	Conteo pre-revisión	Conteo post-revisión
Consulta	666	823
Incidente	619	2099
Requerimiento	8714	5617
Informativo	0	732
No Categorizable	0	728

Tabla 4: Distribución tickets para sub dataset creado

Como se puede apreciar en la tabla, no solo cambiaron las distribuciones para las clases, sino que también se decidió asignar 2 clases nuevas, Informativo y No Categorizable, pues se considerar que estas dos nuevas clases clasifican de mejor forma las situaciones descritas por esos tickets.

Esta situación es relevante para la empresa, ya que, si la información que uno posee es de confianza esto facilita la toma de decisiones, como en el caso de saber específicamente que clientes presentan un mayor grado de incidentes, consultas o requerimientos, ya que ayuda a la hora de decidir qué cantidad de recursos asignar a cada estudio.

Ahora, dentro de las causas de porque existe esta situación en la empresa, se presenta la siguiente lista:

- No existe una capacitación suficiente de los técnicos para poder leer e interpretar de forma correcta la información entregada en los tickets.
- Los técnicos no le dedican el tiempo suficiente al trabajo de clasificación de los tickets, lo que presenta un punto interesante, pues no se ha indicado que existe un bono o beneficio por el número de tickets que se atiendan.
- Tecnolex no cuenta con la infraestructura suficiente para poder controlar el registro de información a sus sistemas.

Se propondrían las siguientes soluciones de forma respectiva para solucionar las posibles causas indicadas:

- Tecnolex debería designar más recursos a la hora de capacitar a sus técnicos de Helpdesk.
- Implementar un mayor control del tiempo designado por sus funcionarios a las distintas actividades que realizan.
- Solicitar a SysAid que implemente una herramienta que permita clasificar los tickets de forma semi-automática, de forma que se presenta una recomendación a los técnicos la cual tendrán que revisar.

Se espera que con este trabajo de memoria se pueda entregar una herramienta a Helpdesk que ayude a los técnicos del área a clasificar los tickets, ya sea haciendo este proceso más rápido, y/o permita entregar una primera opción de la clasificación que podría tomar el ticket, y el técnico podría partir a trabajar desde ese punto.

3.OBJETIVOS

Con el fin de lograr ya sea acabar, o reducir la situación presente en la clasificación de los tickets, se presentan los siguientes objetivos.

3.1. Objetivo General:

Generar un modelo de lenguaje natural que permita sugerir una clasificación de los tickets generados en la mesa de ayuda, con el fin de ayudar a los técnicos del área a clasificar los tickets.

3.2. Objetivos Específicos:

- Identificar las clasificaciones realizadas a los tickets y definir nuevas en caso que las anteriores no sean suficientes.

- Identificar el formato y presentación de los textos indicados por los campos título y descripción de los tickets y estandarizarlos para poder ser trabajados.

- Establecer los algoritmos de lenguaje de procesamiento natural para construir el modelo clasificador, y luego evaluar los resultados entregados por este.

- Presentar recomendaciones para las clasificaciones de los tickets, en base a los resultados obtenidos.

4. MARCO CONCEPTUAL

Se separará el marco conceptual en 4 secciones, en donde la primera corresponde a elementos relacionados al trabajo a realizar, el segundo corresponde a técnicas de procesamiento de texto, la tercera corresponde a los modelos de machine learning que se aplicaran al trabajo y la cuarta corresponde a las métricas de evaluación a utilizar.

4.1. Conceptos Generales:

Para esta parte del informe, las definiciones se basan en un documento entregado por Tecnolex.

4.1.1. **Ticket de soporte técnico:**

El ticket de soporte técnico es el elemento básico de cualquier trabajo relacionado con la experiencia del cliente, lo que le permite a las empresas crear, actualizar y resolver cualquier problema que puedan tener las personas que la contactan. Es un boleto digital generado por un sistema de tickets a partir de las solicitudes entrantes realizadas por los usuarios.

4.1.2. **Incidente**:

Evento que no forma parte del desarrollo habitual del servicio y que causa, o puede causar, una interrupción del mismo o una reducción de la calidad de dicho servicio.

4.1.3. **Consulta**:

Petición de un servicio, la cual incluye la asistencia, el consejo, información, cambio de un estándar simple, aprovisionamiento de consumibles, acceso a un servicio, incluso una queja o felicitación.

4.1.4. **Requisito**:

Proceso que se refiere a todas las solicitudes que generan los usuarios y que no están asociadas a un incidente en un servicio, como por ejemplo un desarrollo de otro proceso, una modificación, habilitación de un acceso, entre otros.

4.2. Procesamiento de texto:

4.2.1. **Tokenización**:

Proceso en que los textos se dividen en token, los cuales típicamente son constituidos por palabras individuales. El objetivo de este proceso es dividir el texto en pequeños fragmentos significativos.

4.2.2. **Limpieza y estandarización:**

Al trabajar con palabras en programación se consideran diferentes aquellas que contengan mayúsculas en vez de minúsculas, por lo que se tiende estandarizar las palabras para finalmente solo usar palabras en minúsculas. En esta parte del proceso también se remueven signos de puntuación, ya que éstos no aportan valor a las conclusiones futuras.

4.2.3. **Stopwords**:

No todas las palabras en un texto proveen de un significado como tal, las palabras que no poseen significado se denominan stopwords. Algunos de los ejemplos comunes de stopwords en español serían "de", "la", "que", "el", "en", y" entre otros. **(KeepCoding, 2023)**

4.2.4. **Stemming:**

Es el proceso de reducir una palabra a su raíz o forma base, eliminando sufijos, prefijos, inflexiones y otras formas modificadas. En general, se utiliza para mejorar el procesamiento de lenguaje natural.

4.2.5. **Lematización**:

Es el proceso de convertir una palabra a su forma base, considerando su significado y su uso en el contexto. Es muy utilizado para reducir la cardinalidad del vocabulario asociado para diferentes formas flexionadas con un único token. **(KeepCoding, 2023)**

4.2.6. **TF-IDF:**

Algoritmo que es el resultado de la multiplicación de otros dos algoritmos. Estos algoritmos son Term Frequency (TF) y Inverse Document Frequency (IDF). El primer algoritmo indica la frecuencia de una palabra en un documento con respecto al largo de este mismo, es

decir, la cantidad de veces que aparece una palabra en un documento dividido por el total de palabras en este. **(Cáceres, 2020)**

$$tf(t,d) = \frac{\textbf{Número de ocurrencias de un termino en el documento}}{\textbf{Número de palabras totales en el documento}}$$

El segundo algoritmo (IDF) evalúa que tan única es una palabra. Para esto se utilizan múltiples documentos y se calcula el cociente entre el número total de documentos y el número de documentos que tienen el término en cuestión.

$$idf(t,d) = log(\frac{\textbf{Número de ocurrencias de un termino en el documento}}{\textbf{Número de documentos que contienen el termino}})$$

4.3. Machine Learning:

Los modelos de machine learning se pueden dividir entre algoritmos supervisados, no supervisados y de refuerzo. Debido a la naturaleza del problema y el requerimiento de Tecnolex, se considera implementar solamente modelos de tipo supervisado. Dentro de este tipo de modelos, existen algunos que permiten realizar clasificación de texto, tales como Regresiones logísticas, SVM, Árboles de Decisión, Algoritmos Probabilísticos (Naive Bayes) y Redes Neuronales, entre otros. En este trabajo se considera hacer uso de una Regresión Logística Multinomial, un modelo SVM y u árbol de decisión, donde la descripción de cada uno de ellos se presenta a continuación.

4.3.1. **Support Vector Machine**:

Algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz. El objetivo del algoritmo SVM es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos. "De la mejor forma posible" implica el hiperplano con el margen más amplio entre las dos clases, representado por los signos más y menos en la siguiente figura. El margen se define como la anchura máxima de la región paralela al hiperplano que no tiene puntos de datos interiores. El algoritmo solo puede encontrar este hiperplano en problemas que permiten separación lineal; en la mayoría de los problemas prácticos, el algoritmo maximiza el margen flexible permitiendo un pequeño número de clasificaciones erróneas. **(mathworks, s.f.)**

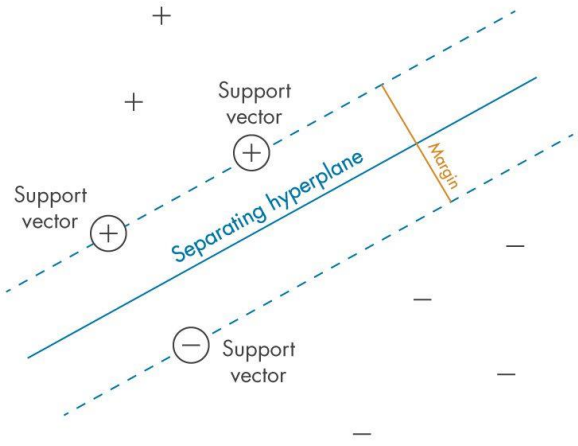

Figura 6: Esquema Support Vector Machine

4.3.2. **Regresión Logística:**

Algoritmo de clasificación que se suele utilizar para conocer la probabilidad de que una variable se encuentra en una categoría. En la regresión logística, la variable se trata de una variable binaria que tiene datos codificados como 1 o como 0.

La principal diferencia con un modelo de regresión lineal es que la regresión logística entrega un resultado constante, mientras que la regresión lineal un resultado continuo, esto porque la primera utiliza el método de máxima verosimilitud y la segunda el de mínimos cuadrados ordinarios.

Los modelos de regresión logística son extensibles a la clasificación de múltiples clases a la vez, en esta variante cada categoría se compara de a una contra el resto de las categorías, es decir estima de manera separada.

Aunque esta estrategia es popular a la hora de clasificar, tiene diversos problemas como, por ejemplo, debido a que los clasificadores solo ven como positivo los datos de su misma clase y negativo todos los demás, hay un desbalance a la hora de entrenar el clasificador que usará el modelo, ya que en el conjunto de datos de entrenamiento existirán muchos más negativos que positivos.

4.3.3. **Árbol de Clasificación (IBM, 2023):**

Un árbol de clasificación es un tipo de árbol de decisiones. Utiliza la medida de impurezas de Gini para clasificar los registros en las categorías del campo objetivo. Las predicciones se basan en combinaciones de valores en los campos de entrada.

Un árbol de clasificación calcula la categoría de destino pronosticada para cada nodo en el árbol. Este tipo de árbol se genera cuando el campo de destino es categórico.

Cada nodo se divide en dos o más nodos hijo para reducir el valor de impureza Gini del nodo. La impureza de Gini es una función que penaliza más incluso la distribución de valores objetivo y se basa en las estadísticas de frecuencia de destino y en el número de filas de datos correspondientes al nodo. Los nodos hijo correspondientes a las categorías de predictores dados se fusionan cuando el incremento correspondiente en la impureza de Gini es tolerable dentro del límite especificado. Para cada nodo, el predictor que reduce el valor de impurezas de Gini más se selecciona para la división del nodo.

El proceso de creación de un árbol de decisiones se inicia con el nodo raíz que corresponde a todas las filas de los datos. Cualquier nodo se divide en nodos hijo hasta que no es posible ninguna mejora adicional en la impureza de Gini, o el número de filas de datos correspondientes al nodo se vuelve demasiado pequeño. El proceso también se detiene si el número de nodos en el árbol de decisiones pasa a ser demasiado grande.

La potencia predictiva que se notifica para un árbol de clasificación es el recuento ajustado R^2. Se obtiene calculando la precisión de clasificación de árbol sobre el modelo constante y dividiéndolo por el error de clasificación de modelo constante. El modelo constante siempre predice la modalidad de destino y su precisión de clasificación se estima según la frecuencia de la modalidad. Se informa de un árbol de clasificación predictiva fiable cuando su fuerza predictiva es mayor que un umbral predeterminado del 10%.

4.4. Métricas de Evaluación (Cáceres, 2020):

4.4.1. **Logarithmic Loss**:

Es una métrica de comparación de modelos que está basada en la función de verosimilitud. Los valores que entrega esta métrica no pueden interpretarse por sí solos, aunque siempre un menor valor implica un mejor modelo.

4.4.2. **Matriz de Confusión:**

La matriz de confusión es una tabla que se usa para poder medir el desempeño de un modelo de clasificación. Esta tabla, se divide en filas según la etiqueta real y en columnas en función del valor predicho.

Los elementos dentro de una matriz de confusión son los siguientes:

1- Verdaderos positivos (VP): Número de casos en los que el modelo predijo que pertenecía a la clase y así era.
2- Falsos positivos (FP): Número de casos en que se predice que pertenece a la clase, pero realmente no era así.
3- Verdaderos negativos (VN): Número de casos en que no pertenece a la clase y el modelo efectivamente predice eso.
4- Falsos negativos (FN): Número de casos en que se predice que no pertenece a la clase, cuando si pertenecía.

Algunas métricas que derivan de la matriz de confusión son:

1- Accuracy: Mide con qué frecuencia un clasificador está correcto.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{VP} + \boldsymbol{VN}}{\boldsymbol{VP} + \boldsymbol{Vn} + \boldsymbol{FP} + \boldsymbol{FN}}$$

2- Tasa de error: Mide con qué frecuencia un clasificador está correcto

$$\boldsymbol{Tasa\ de\ error} = \frac{\boldsymbol{FP} + \boldsymbol{FN}}{\boldsymbol{VP} + \boldsymbol{VN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

3- Ratio de verdaderos positivos: Mide cuantas de las veces en que un registro pertenecía a la clase el modelo predijo exactamente eso.

$$\boldsymbol{Ratio\ de\ verdaderos\ positivos} = \frac{\boldsymbol{VP}}{\boldsymbol{VP} + \boldsymbol{FN}}$$

4- Ratio de verdaderos negativos: Mide cuantas de las veces en que un registro no pertenecía a la clase el modelo predijo exactamente eso.

$$\boldsymbol{Ratio\ de\ verdaderos\ negativos} = \frac{\boldsymbol{VN}}{\boldsymbol{VN + FP}}$$

5- Ratio de falso positivos: Mide cuantas de las veces en donde no pertenece a la clase, el modelo predice que sí.

$$\boldsymbol{Ratio\ de\ falsos\ positivos} = \frac{\boldsymbol{FP}}{\boldsymbol{FP + VN}}$$

6- Precision: Mide cuantas veces de las que el modelo predice que si, efectivamente era así

$$\boldsymbol{Precision} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FP}}$$

7- Prevalence: Mide que tanto la condición de pertenecer a la clase ocurre en la muestra.

$$\boldsymbol{Prevalence} = \frac{\boldsymbol{VP + FN}}{\boldsymbol{VP + VN + FP + FN}}$$

4.4.3 Área bajo la curva:

La curva ROC es una forma de mostrar el desempeño de un clasificador binario de manera visual.

Los algoritmos de clasificación como la regresión logística no solo son capaces de discernir la clase a la que pertenece el registro, sino que también pueden entregar la probabilidad con la que esto pasa.

Se le denomina "threshold" a la probabilidad mínima con la que un registro será admitido en la clase, comúnmente se utiliza un threshold de 0.5 lo que significa que todos aquellos registros que tengan una probabilidad igual o mayor a 0.5 serán clasificados como parte de la clase.

La curva ROC, representa la tasa de verdaderos positivos en función de la tasa de falsos positivos para cada valor posible de

threshold. Mientras mayor sea el área bajo la curva, mejor será la performance del clasificador.

4.4.4 F1-Score:

La métrica F1 score es una función en base a la ratio de verdaderos positivos (recall) y la precisión (precisión) del clasificador, su uso se recomienda cuando se tiene un problema de clasificación con clases desbalanceadas.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

5. METODOLOGÍA

Para este trabajo se hará uso del proceso KDD o "Knowledge Discovery in Databases" propuesto por parte de Fayyad, Piatetsky-Shapiro y Smyth en 1996 es definido como **(Fayyad, 1996)**:

"El proceso no trivial de identificar patrones en los datos que resulten válidos, originales, potencialmente útiles y finalmente comprensibles"

Este corresponde a una serie de etapas y pasos realizados con el fin de descubrir patrones en los datos que puedan aportar inteligencia para la resolución de un problema de negocio.

Los pasos más importantes de este proceso son la selección y preparación de los datos, el pre procesamiento de los datos, la transformación de los datos, minería de datos y evaluación e interpretación de los resultados.

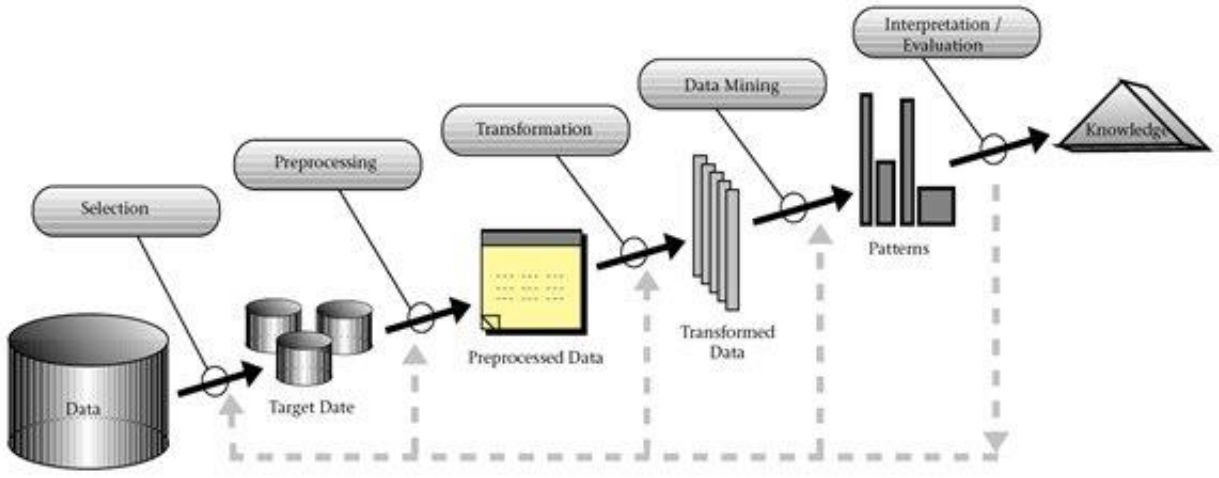

Figura 7: Representación gráfica del proceso KDD

Para la metodología de este trabajo se tomó en cuenta las siguientes consideraciones para cada uno de los puntos del proceso KDD:

1. Selección y preparación de datos: Los datos fueron entregados por parte de Tecnolex en una planilla de Excel. A pesar de que se cuenta con la información de todos los atributos de los tickets, solo se trabajó con los atributos que correspondían a los textos presentes en los tickets y su clasificación. Para esta base de datos se cuenta con 116.835 y 23 columnas en la planilla Excel entregada.

2. Pre-procesamiento de los datos: Dentro de los atributos Categoría y Tipo de Requerimiento se debe estudiar las distribuciones de datos de sus clases, y para el atributo de Categoría este cuenta con 14 clases distintas, por lo que se debió estudiar cómo trabajar en clasificar una variable con esa cantidad de posibles valores a tomar.

3. Transformación de los datos: Dentro de los textos entregados, se encontraban palabras en inglés, por lo que se debió traducir estas palabras a español de forma de que el modelo propuesto lograra clasificar. Tambien la información de los títulos y descripciones de los tickets se encuentra en variables distintas.

4. Data Mining: Se debió considerar el tamaño de la base de datos y la naturaleza de los datos, como por ejemplo una variable con 14 posibles clases distintas, a la hora de elegir que modelos utilizar para clasificar los datos.

5. Evaluación de los resultados: Se tomó en cuenta la naturaleza de los datos y la frecuencia con la que se repiten ciertos eventos para el área de Helpdesk a la hora de evaluar los resultados.

6. ALCANCES

- Con respecto a la data a entregar, se considerarán tickets hasta los últimos 2 años.

- Los tickets pueden ser generados por cualquiera de los clientes o Tecnolex o por uno de sus funcionarios, indicado por el atributo Estudio de los tickets.

- Para la clasificación de los tickets, estos tienen 4 niveles de clasificación, en donde estos niveles corresponden a tipo de requerimiento, primera categoría, sub categoría y tercera categoría. Para este trabajo se espera conseguir clasificar el tipo de requerimiento y su primera categoría.

- Solo se realizará una clasificación de los tickets, y una vez esta haya sido entregada junto con los modelos, no se realizará una nueva en caso que se hayan generado nuevos tickets durante la realización de este trabajo.

7. RESULTADOS ESPERADOS

En base al avance del trabajo, los resultados esperados se pueden dividir en dos puntos principales:

- El primer resultado es la entrega del modelo de clasificación junto con indicaciones de cómo usarlo.

- El segundo resultado es la entrega de un reporte con las clasificaciones realizadas a los tickets entregados, además de cualquier descubrimiento extra que se haya encontrado al estudiar la data.

8. Desarrollo Metodológico

Tomando como base la metodología de trabajo decidida, la situación presentada por Tecnolex, y otros trabajos en temas similares investigados, se presenta a continuación cada una de las actividades que han sido realizadas para cada una de las etapas del proceso KDD. Se debe indicar que todo el trabajo de programación se realizará con el lenguaje Python, pero parte del procesamiento y limpieza de los datos se ha realizado con Excel y Google Sheets.

8.1 Selección y preparación de los datos

Tecnolex cuenta con servidores en donde se guarda la data correspondiente a los tickets. Se exportaron los datos en formato Excel, los cuales ahora están guardados en el computador personal del memorista. La data tiene una antigüedad máxima de dos años. La data se guardó en un dataframe, el cual contiene 116.835 filas y 23 columnas, siendo las columnas los atributos de los tickets indicados previamente y las filas cada uno de los tickets.

Primero, se desea conocer la cantidad de valores nulos existentes en la base. Al revisar la base de datos, se encontraron 54.814 tickets que contaban con al menos un atributo nulo. En el siguiente gráfico se puede apreciar el porcentaje de datos nulos para cada atributo respecto al total de tickets presente en la base de datos.

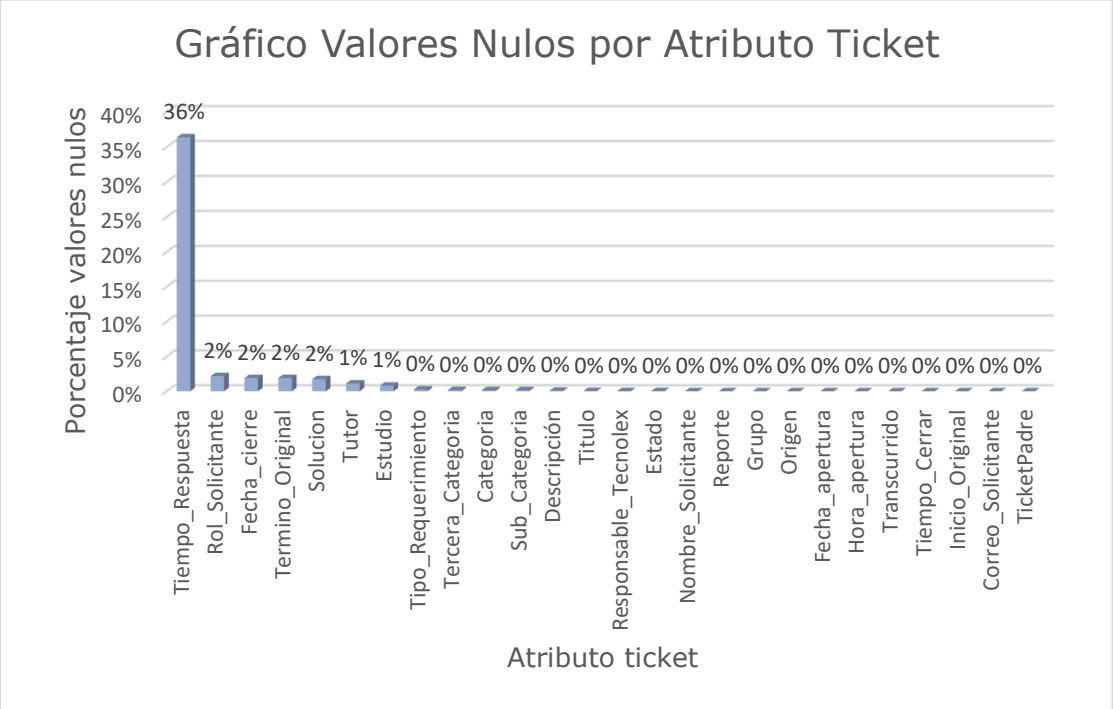

Figura 8: Gráfico de distribución de valores nulos por atributos de los tickets

Como se puede apreciar en el gráfico, la mayoría de los atributos no presenta valores nulos, siendo el caso extremo el tiempo de primera respuesta a los clientes por parte de los técnicos de Helpdesk. El atributo Tiempo_Respuesta corresponde al tiempo que le toma al técnico realizar la primera respuesta al cliente, pero puede ocurrir el caso en que el ticket ingrese a sistema cerrado, por lo que el sistema no ingresaría ningún valor a ese atributo. Aun asi, Tiempo_Respuesta corresponde al grupo de tickets que deberían ser llenados de forma automática por el sistema, lo que da a entender que se debería revisar esta funcionalidad de SysAid con el fin de implementar una señal que diferencie los tickets que han sido ingresado a la base en estado cerrado de los tickets a los cuales todavía no se ha realizado una primera instancia de interacción con el cliente.

8.2 Pre-procesamiento de los datos

Ya que el fin de este trabajo es lograr clasificar los tickets en base a sus variables Tipo Requerimiento y Categoría, se identifican las distribuciones actuales de estos atributos, las cuales se pueden apreciar en los siguientes gráficos. Se presenta la proporción porcentual para cada una de las clases respecto al total de tickets en la base de datos.

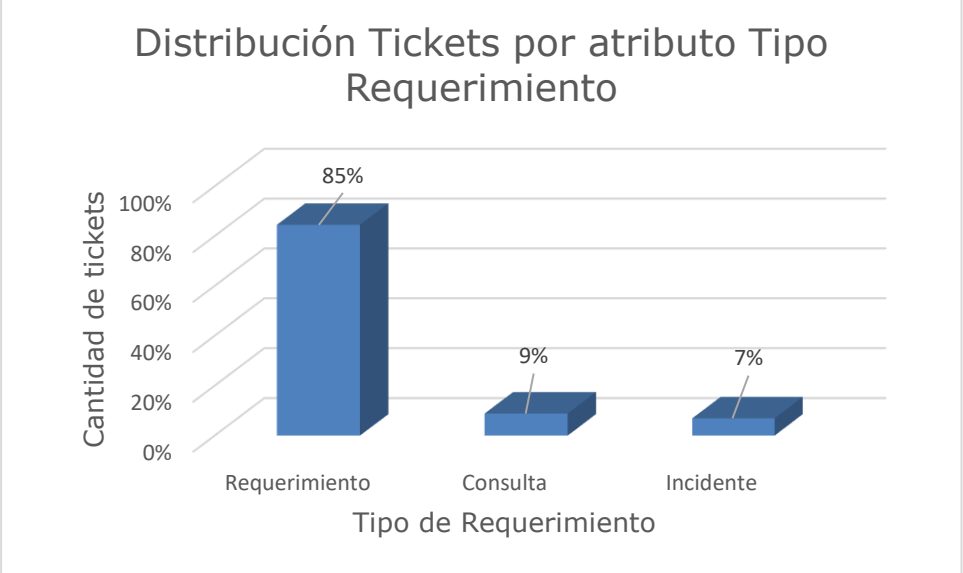

Figura 9: Gráfico de distribución de los tickets por clase para el atributo Tipo Requerimiento

Como se puede apreciar en el gráfico, la mayoría de los tickets se encuentran clasificados como Requerimiento, los cuales corresponden al 85% del total de datos presentes en la base.

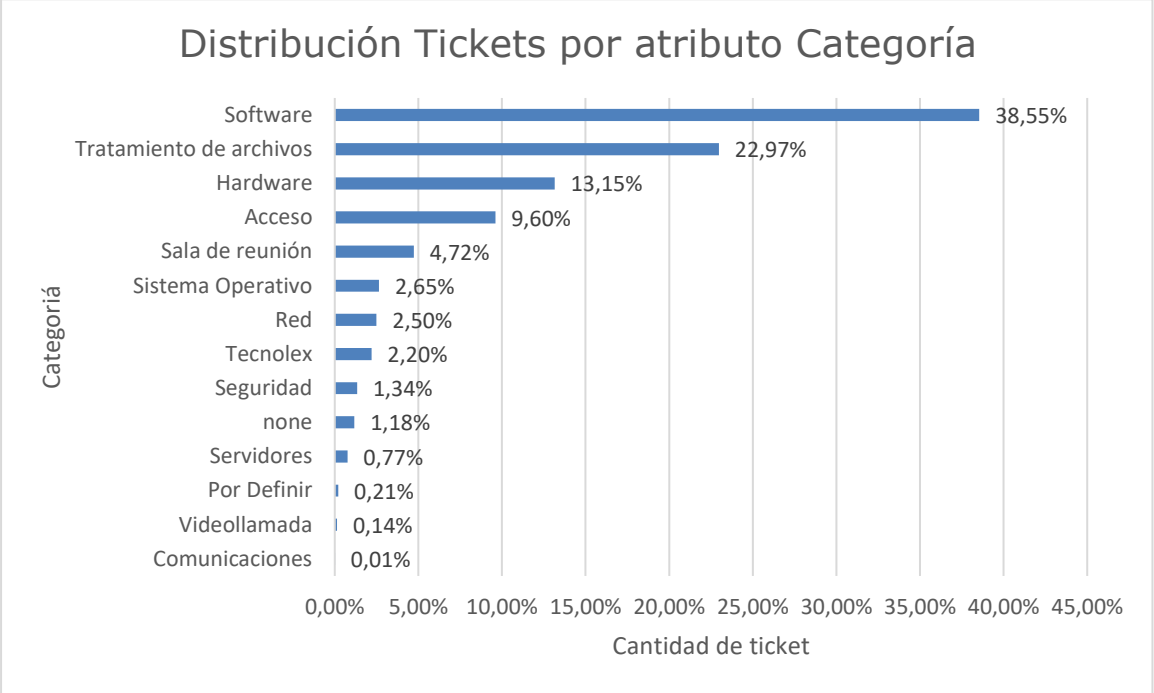

Figura 10: Distribución de los tickets por clase para el atributo Categoría

Mientras que en este gráfico se puede apreciar que los tickets presentan una mayoría a ser clasificados como Software y Tratamiento de Archivos, a la vez que se puede observar que otras categorías cuentan con una muy baja cantidad de datos. También se desea notar que existen tickets clasificados como none y Por Definir por parte de los técnicos del área, por lo que se recomendaría volver a revisar esos tickets, ya sea para asignarlos a categorías ya existentes o definir una nueva para la cual sean considerados apropiados.

A continuación, se decidió identificar las distribuciones de los tickets por su origen, ya que como se indicó previamente el tipo de origen afecta la forma en que se van llenando los atributos de los tickets.

En el siguiente gráfico se puede apreciar las distribuciones de los tickets por su origen.

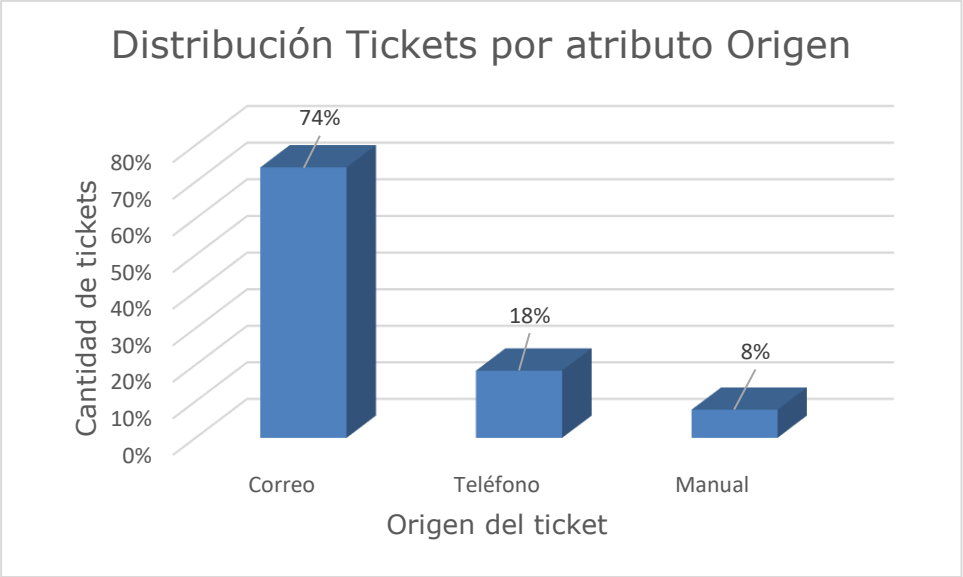

Figura 11: Gráfico distribución de los tickets por su atributo Origen

Como se puede ver, la gran mayoría de los tickets tienden a ser generados por correo, por lo que la mayoría de sus atributos son llenados de forma automática, siendo además el caso que el título y descripción de estos tickets corresponden al asunto y cuerpo del correo respectivamente.

Otro de los puntos importantes de esta etapa fue verificar la declaración realizada por uno de los jefes del área, quien indicaba que las ultimas clasificaciones realizadas estaban incorrectas y no era posible realizar nuevos reportes debido a esto. Para poder verificar la declaración se creó un sub set de 10.000 tickets, y se revisó de forma manual cada uno con el fin de identificar el porcentaje de tickets que estaban bien clasificados para la variable de Tipo Requerimiento. Uno de los puntos importantes que se descubrió revisando la data es que las clasificaciones realizadas previamente por parte de técnicos no son suficiente para todas las posibles situaciones descritas en los tickets, por lo que se propone agregar 2 más que corresponden a Informativo y No Categorizable, ya que se encontraron tickets que solo entregaban información ya sea a un cliente o empleado de Tecnolex, y "No Categorizable" ya que se presentan tickets con una sola palabra y ningún otro contexto, por lo que no cuentan con la suficiente información para poder ser clasificados. Se propone esta segunda clasificación con la idea de dar una alerta a los técnicos de que deberían la forma en la que están ingresando la data a los tickets, ya que no se está anotando suficiente información al ingresar los tickets a la base.

A continuación, se presenta el gráfico con las distribuciones de Tipo Requerimiento para esta muestra de 10.000 tickets, antes y después de haber realizado la revisión manual.

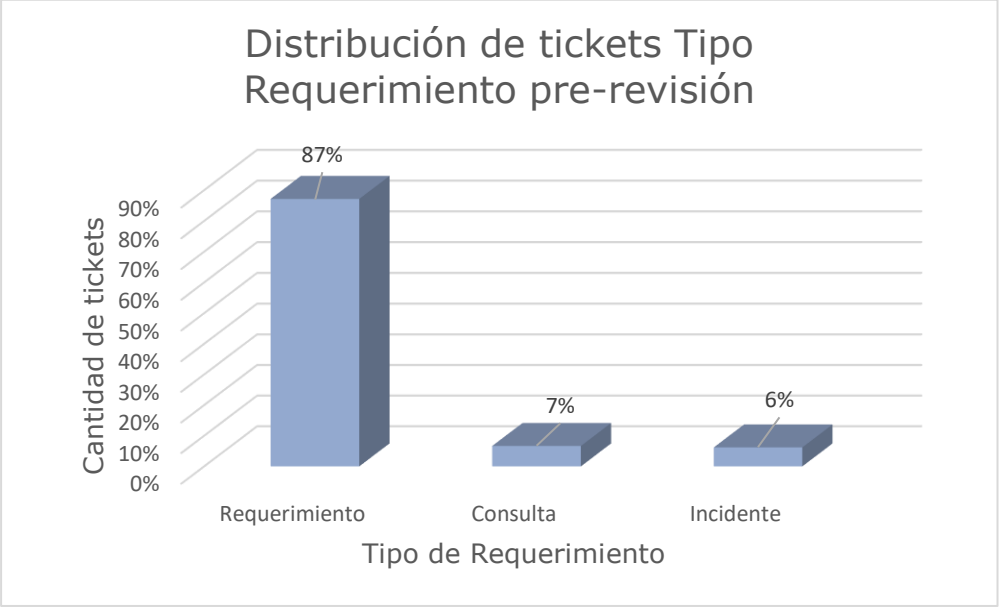

Figura 12: Grafico distribución de los tickets por atributo Tipo Requerimiento para sub set sin revisar

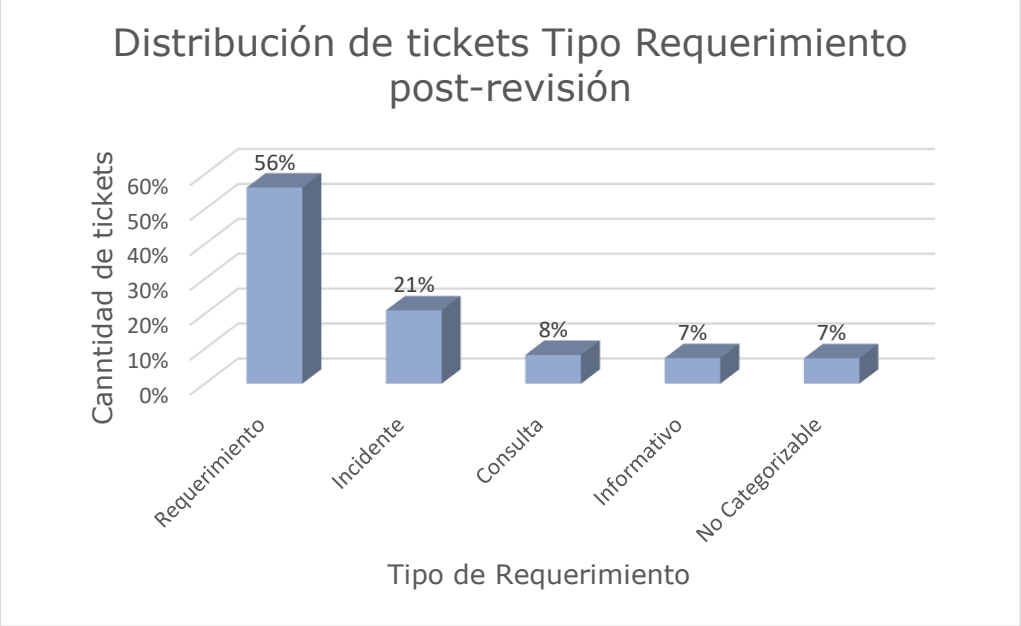

Figura 13: Gráfico distribución de los tickets por atributo Tipo Requerimiento para sub set revisado

Al revisar los gráficos, se puede ver que en ambos se presenta una tendencia a clasificar los tickets con la clase "Requerimiento", pero también se puede ver que después de la revisión este porcentaje disminuyó, mientras que el porcentaje para la clase "Incidente" aumento en gran medida, y se puede

ver que tanto la clase “Informativo” como “No Categorizable” cuentan con un 7% de los datos respectivamente.

La revisión de la declaración realizada no solo se hizo con el atributo “Tipo Requerimiento”, sino que tambien se trabajó con el atributo “Categoría”, pero para este caso se decidió crear otro sub set más, pero esta vez de solo 3.000 tickets elegidos de forma aleatoria, al cual también se le realizó una revisión manual.

A continuación, se muestran las distribuciones obtenidas antes y después de la revisión.

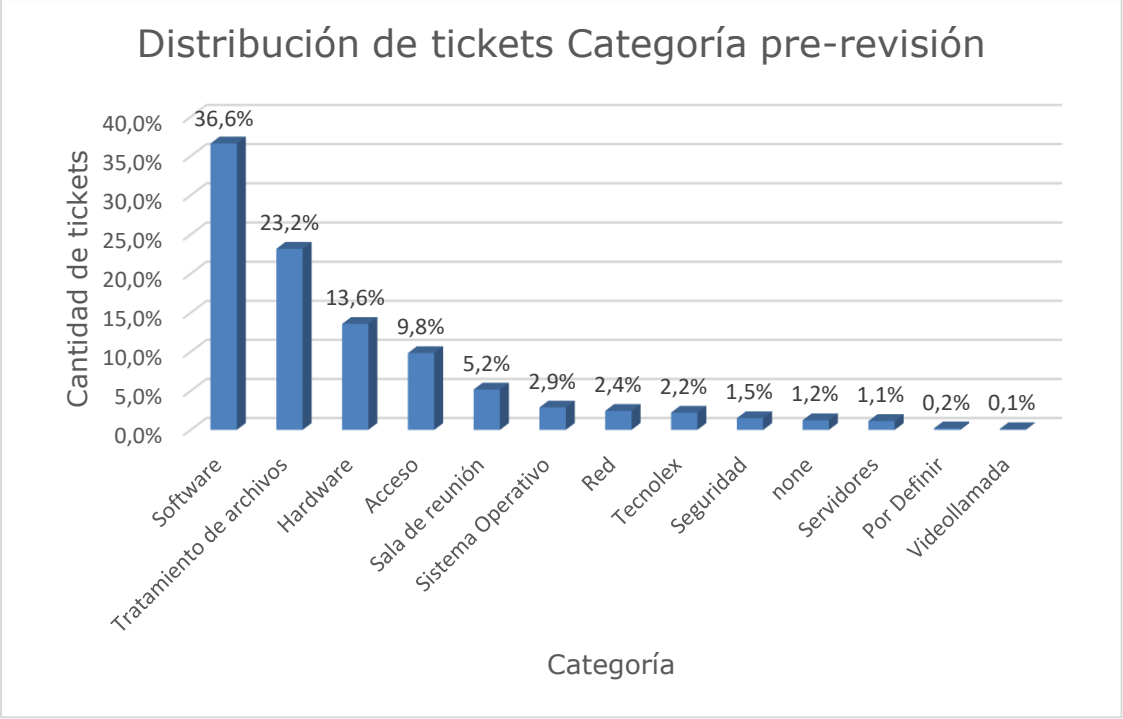

Figura 14: Gráfico distribución de los tickets por atributo Categoría para un sub set sin revisar

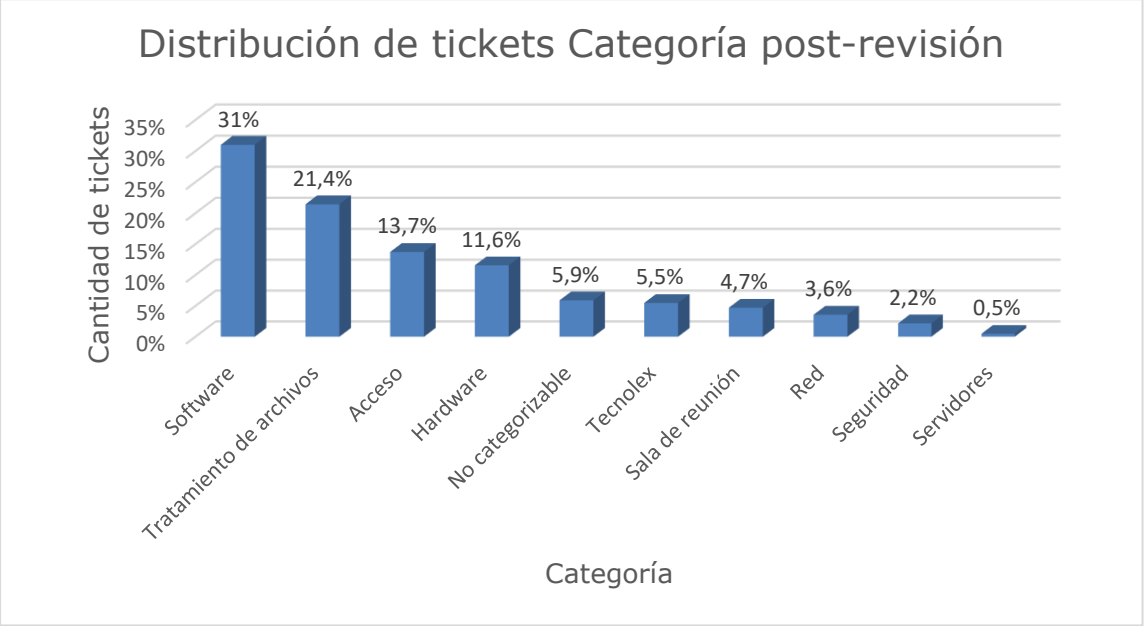

Figura 15: Gráfico distribución de los tickets por atributo Categoría para un sub set revisados

Como se puede apreciar en los gráficos, a pesar de que para este sub set de datos también se presenta un desbalance hacia las clases Software y Tratamiento de archivos, se pueden apreciar diferencias entre las distribuciones, ya que existen ciertas clases como Sistema Operativo o Video llamada, las cuales se ha decidido agregar a otras. Tambien se puede notar que en el primer gráfico existen las clases none y Por definir, en donde se ha decidido clasificar sus datos en otras clases, como se puede apreciar en el segundo gráfico, pero tambien se quiere hacer notar que se ha creado una nueva clase, llamada No categorizable, al igual como se hizo con el atributo Tipo Requerimiento, con el fin de asignar en esta clase todos los tickets que no cuenten con la información suficiente o su información no sea la suficientemente clara para poder ser clasificada.

Tambien se quiso estudiar como distribuían los tickets a través del tiempo, con el fin de identificar si ocurrió algún evento durante el periodo que se está trabajando que podría explicar el desbalance que se encuentra presente en la base.

Primero se va a mostrar las distribuciones a través del tiempo para el atributo Tipo Requerimiento.

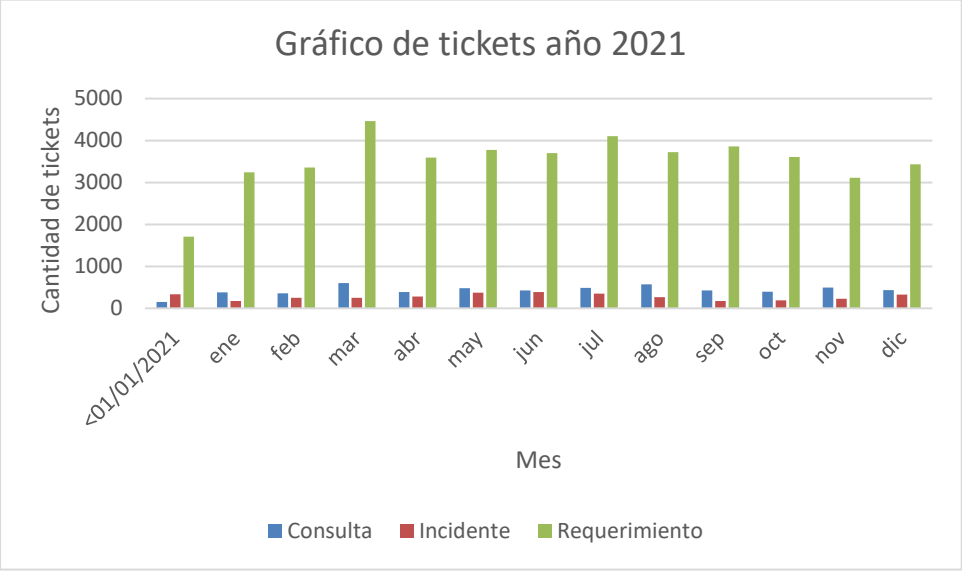

Figura 16: Gráfico distribución de los tickets para el atributo Tipo Requerimiento para el año 2021

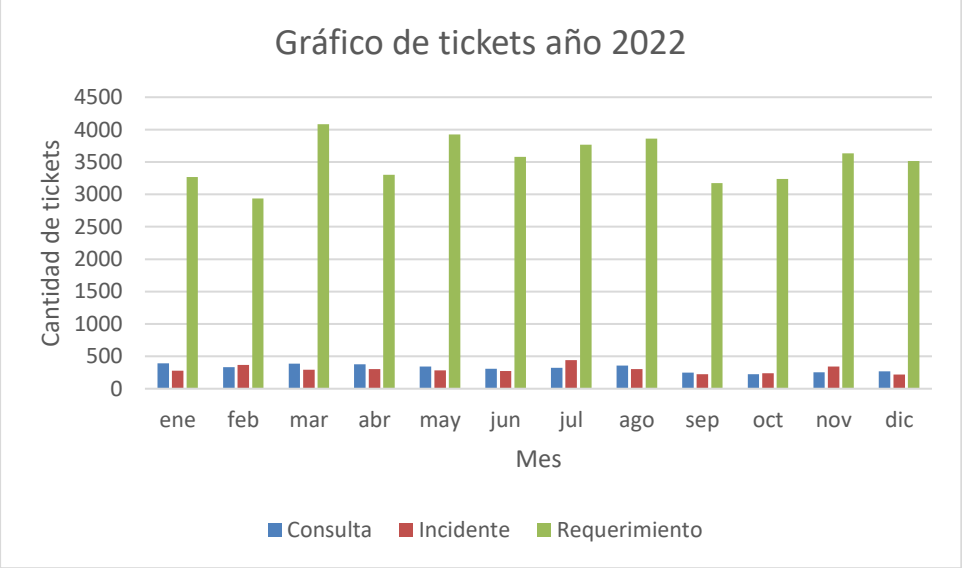

Figura 17: Gráfico distribución de los tickets para el atributo Tipo Requerimiento para el año 2022

Al revisar los gráficos se puede ver que la cantidad de tickets generados entre cada año es distinta, pero las distribuciones se mantienen similares durante el año, siempre manteniendo la desigualdad hacia el valor Requerimiento.

En la siguiente tabla se puede ver el número de tickets para cada una de las clases del atributo Categoría tanto por mes como por año:

Categoría	Año	01	02	03	04	05	06	07	08	09	10	11	12
Acceso	2021	366	365	501	388	325	390	687	382	433	333	336	435
	2022	326	369	419	407	452	365	499	466	392	344	407	397
Comunicaciones	2021	0	0	0	0	0	0	0	1	6	3	0	0
	2022	3	1	2	0	0	0	0	0	0	0	0	0
Hardware	2021	459	494	744	355	590	544	710	680	747	651	542	484
	2022	465	473	549	450	625	554	666	632	530	532	572	481
None	2021	0	1	0	1	0	0	1	21	30	69	19	31
	2022	18	26	42	8	42	11	38	28	14	12	36	29
Por Definir	2021	11	40	48	19	38	23	35	7	8	11	0	0
	2022	0	0	0	0	0	0	0	0	0	0	0	0
Red	2021	84	150	166	82	137	159	106	109	110	76	86	80
	2022	75	90	119	80	84	79	79	83	51	95	104	82
Sala de reunión	2021	158	121	280	167	184	145	157	185	199	190	225	228
	2022	172	85	219	223	296	275	221	269	191	184	278	221
Seguridad	2021	0	0	0	60	137	76	126	56	37	34	56	85
	2022	97	137	77	112	61	26	46	57	30	25	48	20
Servidores	2021	42	67	150	165	36	27	37	29	28	16	21	20
	2022	20	18	21	23	32	13	8	16	15	5	19	14
Sistema Operativo	2021	139	142	146	183	122	140	135	119	136	101	89	99
	2022	111	93	133	123	97	69	87	79	153	81	130	108
Software	2021	1503	1565	1989	1666	1616	1796	1736	1679	1594	1525	1375	1543
	2022	1526	1463	1901	1467	1648	1624	1739	1805	1435	1545	1692	1804
Tecnolex	2021	87	102	179	117	147	118	101	84	102	85	56	83
	2022	72	89	127	96	90	96	75	79	85	75	71	61
Tratamiento de archivos	2021	952	953	1161	1075	1308	1099	1111	1209	1029	1094	1026	1106
	2022	1047	787	1159	991	1116	1041	1079	1010	748	792	850	760
Videollamada	2021	0	0	0	0	0	0	0	0	0	11	8	11
	2022	8	13	5	4	4	8	1	4	2	12	22	23

Tabla 5: Tabla comparativa número de tickets para cada clase de categoría por mes y por año

Como se puede ver en la tabla, los tickets mantienen distribuciones similares para este atributo a través de los años.

Luego, se decidió unir las columnas Titulo y Descripción en una sola nueva columna llamada Texto, ya que ocurría el caso en que la variable Descripción contaba con valor nulo pero el Titulo lograba explicar la situación lo suficiente para poder ser clasificado.

Por último, como se ha indicado previamente en esta sección del informe, se ha descubierto que existe un gran desbalance en la base de datos, tanto para el atributo Tipo Requerimiento como para Categoría. Es por esto que se ha decidido crear dos dataset nuevos en los cuales se han balanceado los datos.

En la base inicial se puede ver que el valor Incidente es el que cuenta con la menor cantidad de datos, por lo que se decidió eliminar de forma

aleatoria datos para la clase Consulta y Requerimiento, con el fin de que todas las clases tengan la misma cantidad de tickets.

Para el caso del atributo Categoría se implementó otra técnica, se decidió agrupar las clases existentes que tuvieran relación entre ellas en nuevas clases más grandes, y de esta forma hacer que el desbalance fuera más pequeño.

Por lo tanto, para este trabajo de modelaje se va a estar contando con 5 data sets distintos, en donde cada uno se encuentra guardado en una planilla de Excel distinta:

- El primer dataset corresponde al original con todos los tickets entregados por parte de Tecnolex.
- El segundo dataset está conformado por los 10.000 tickets más recientes, los cuales han sido revisados de forma manual para el atributo Tipo Requerimiento. Se creó este segundo data set para descubrir el porcentaje de tickets que han sido incorrectamente clasificados por parte de los técnicos.
- El tercer dataset cuenta con 3.000 tickets, los cuales han sido de forma aleatoria para revisar el atributo Categoría. Tambien se creó este data set más pequeño para tomar menos tiempo al probar los modelos.
- El cuarto dataset se formó al eliminar de forma aleatoria tickets para la clase Consulta y Requerimiento hasta que las 3 clases correspondientes a Tipo Requerimiento tuvieran la misma cantidad de valores. De esta forma se deseó estudiar cómo se comportan los modelos al trabajar con clases con la misma cantidad de tickets.
- Y el quinto dataset posee la misma cantidad de tickets que el original, pero ahora todas las clases del atributo Categoría han sido agrupados en nuevas clases más generales. Estas nuevas corresponden a las indicadas en las siguientes listas:

    - Administración y Comunicación
    - Hardware y Redes
    - Seguridad y Acceso
    - Software
    - Tratamiento de Archivos

8.3 Transformación de los datos

Antes de crear un tokenizador para poder aplicar los modelos de Machine Learning, se eliminó de los tickets con origen igual a correo las firmas de las empresas, avisos de información confidencial, avisos de correos externos, información del emisor y receptor de los correos y finalmente los logos de las empresas. Luego, al encontrarse texto en inglés, se exporto la base a Google Sheets, en donde haciendo uso de la función GoogleTranslate se tradujeron las palabras de inglés a español.

A continuación, se creó el tokenizador, el cual cuenta con las siguientes funciones:

- Tokenizar los textos
- Eliminar símbolos de puntuación excepto underscore
- Eliminar caracteres que contienen letras y números
- Eliminar caracteres numéricos
- Pasar las letras a minúsculas
- Eliminar stopwords en español cuando se le indica
- Lematizar los textos cuando se indica
- Aplicar stemming cuando se indica

Se presenta un ejemplo de cómo quedaría un texto aplicando las distintas funciones del tokenizador:

Aplicación del Tokenizador	Texto
Texto Original	RV: Por favor Revisar Urgente/ RV: para su acción: Premio Draft para la Revisión 720BH A23GR00092 WV Chile Estimados, favor. Po dría Transformar El Archivo adjunto una pal abra?
Texto Tokenizado	['rv', 'por', 'favor', 'revisar', 'urgente', 'rv', ' para', 'su', 'acción', 'premio', 'draft', 'para', ' la', 'revisión', 'wv', 'chile', 'estimados', 'favo r', 'podría', 'transformar', 'el', 'archivo', 'adj unto', 'una', 'palabra']
Texto Tokenizado sin Stopwords	['favor', 'revisar', 'urgente', 'acción', 'premio ', 'draft', 'revisión', 'chile', 'estimados', 'favo r', 'transformar', 'archivo', 'adjunto', 'palabr a']
Texto Tokenizado con Lematización	['rv', 'por', 'favor', 'revisar', 'urgente', 'rv', ' para', 'su', 'acción', 'premio', 'draft', 'para', ' el', 'revisión', 'wv', 'chile', 'estimado', 'favor' , 'poder', 'transformar', 'el', 'archivo', 'adjun to', 'uno', 'palabra']
Texto Tokenizado con Stemming	['rv', 'por', 'favor', 'revisar', 'urgent', 'rv', 'p ara', 'su', 'acción', 'premio', 'draft', 'para', 'l a', 'revisión', 'wv', 'chile', 'estimado', 'favor', 'podría', 'transformar', 'el', 'archivo', 'adjunt o', 'una', 'palabra']

Tabla 2: Ejemplo aplicación de tokenizador a texto

Para este trabajo se eliminaron los stopwords y se aplica lematización.

Finalmente, se hace uso del algoritmo TF-IDF para lograr llevar los textos a su respectiva representación vertical.

8.4. Data Mining

Con el fin de lograr entregar una sugerencia para las clasificaciones de los tickets, tanto para el atributo Categoría como Tipo Requerimiento se ha hecho uso de 3 modelos de clasificación distintos: Regresión Logística, SVM y un árbol de clasificación.

Se eligieron estos 3 tipos de modelos conocidos ya que se consideró que serían apropiados para poder responder las siguientes 4 preguntas:

- Efecto de contar con clasificaciones con grandes cantidades de clases a la hora de querer clasificar.
- La naturaleza de los datos, ya que, al corresponder a tickets para una mesa de ayuda de servicios informáticos, dentro de los textos a trabajar se encuentra una gran cantidad de palabras técnicas relacionadas a software y hardware, por lo que se desea conocer el efecto de estas palabras a la hora de clasificar.
- El efecto de contar con más de 100.000 tickets, por lo que se desea saber cómo los modelos se comportan con este volumen de datos, la relación entre los atributos a estudiar y los textos presentes en los tickets, ya que se desea comparar los resultados al aplicar modelos lineales y no lineales.
- Y finalmente, como se ha indicado previamente, no se tiene la certeza de que todos los datos para entregar se encuentren correctamente clasificados, inclusive, como se ha visto al comparar las distribuciones para los diferentes dataset de estudio que se han creado, algunos tickets han sido asignados a clases incorrectas, por lo que se desea estudiar el efecto de este fenómeno sobre las clasificaciones realizadas por los modelos.

Y esta elección se tomó basada en las siguientes razones:

- Regresión Logística: Sirve como un buen modelo base como punto de partida para estudiar cómo se comportan los datos al clasificarlos, a la vez que es un modelo fácil de interpretar y pide un menor requerimiento de recursos comparado con otros modelos a usar, lo que ayuda a la hora de querer estudiar las clasificaciones de los tickets en los distintos datasets creados.
- Máquinas de Vectores de Soporte(SVM): Presenta buenos resultados al trabajar con variables multiclase, son resistentes al sobre-ajuste y a base de datos desbalanceadas, y permite trabajar con datos no lineales, lo que corresponde a este trabajo.

- Árbol de clasificación: Son más eficientes a la hora de trabajar con datos mixtos y atípicos, a la vez de ser fáciles de interpretar, pero a diferencia de las SVM no presenta resistencia al sobreajuste y a la generalización de nuevos datos.

Para correr los modelos se ha decidido separar los dataset en un grupo para entrenamiento correspondiente al 80% de los datos y un grupo para test igual al 20% restante.

Dataset	Atributo	Modelo		
		Regresión	SVM	Árbol
1ro	Tipo Requerimiento	X	X	X
	Categoría	X	X	X
2do	Tipo Requerimiento	X	X	X
	Categoría			
3ro	Tipo Requerimiento			
	Categoría	X	X	X
4to	Tipo Requerimiento	X	X	X
	Categoría			
5to	Tipo Requerimiento			
	Categoría	X	X	X

Tabla 6: Uso de los modelos para cada atributo en cada una de las bases

Como se puede apreciar en la tabla, se puede observar que modelos se corrieron sobre que dataset y para estudiar cual atributo.

Para la primera base de datos indicada se corrieron los 3 modelos elegidos tanto para el atributo Tipo Requerimiento como para la Categoría.

Para el segundo dataset solo se corrieron los 3 modelos para el atributo Tipo Requerimiento, ya que se deseaba conocer cómo se comportan los modelos con datos que han sido revisados y se sabe que están bien clasificados.

Mientras tanto para el tercer dataset también se corrieron los 3 modelos, pero solamente para el atributo Categoría, con el fin de realizar el mismo estudio que para el dataset anterior, pero esta vez con un grupo de datos elegidos de forma aleatoria.

En el cuarto dataset también solo se corrieron los modelos para el atributo Tipo Requerimiento, ya que se deseaba conocer que efecto tendría sobre los modelos el haber realizado el balanceo de las clases.

Finalmente, para el quinto dataset se desea conocer el impacto de haber creado nuevas clases que permitieran balancear las clases sobre los modelos elegidos.

8.5. Análisis e interpretación de resultados

Para poder estudiar los modelos utilizados en este trabajo, se hizo uso de dos herramientas principales: Se utilizó un classification report compuesto por la precisión del modelo, su recall, su F1-Score y el Support, y además se creó una matriz de confusión normalizada. Se eligieron estas métricas debido a que las clases para ambos atributos iniciaron desbalanceadas, hasta que se aplicaron técnicas para balancearlas, por lo que se desea conocer el efecto de haber balanceado las bases sobre los modelos.

8.5.1. Primer dataset

Para el primer dataset con el que se trabajó, se corrieron los 3 modelos y se logró las métricas que se pueden apreciar en el Anexo A y C de este informe.

En ellas se puede apreciar que en los 3 modelos la clase Requerimiento presenta valores mucho más grandes comparado con el resto de las clases de este atributo, lo que se puede explicar por el gran desbalance presente en el dataset. Esto se puede visualizar en las matrices presentes en el Anexo B de este informe, en donde se puede apreciar que los modelos tienden a clasificar todos los tickets como Requerimiento. A la vez, al revisar los resultados de los modelos al intentar clasificar las clases del atributo Categoría, se puede apreciar que los modelos presentan resultados bastante disparejos, en donde las métricas para algunas presentan valores cercanos a 1, mientras que para otras clases sus métricas tienden a ser nulas. Este se puede explicar por el gran número de clases que deben intentar clasificar los modelos a la vez, y también por el gran desbalance de datos entre clases. Esto se puede apreciar de forma gráfica en las matrices de confusión para los modelos presentes en el Anexo D de este informe.

En la siguiente tabla se puede observar la precisión obtenida para los distintos modelos dependiendo de la clase de los atributos.

Atributo	Clase	Regresión	SVM	Árbol
Tipo Requerimiento	Consulta	0.55	0.52	0.40
	Incidente	0.64	0.67	0.48
	Requerimiento	0.88	0.88	0.89
Categoría	Acceso	0.73	0.69	0.65
	Comunicaciones	0.00	0.00	0.00
	Hardware	0.72	0.70	0.65
	Por Definir	0.30	0.26	0.25
	Red	0.74	0.72	0.64
	Sala de Reunión	0.87	0.86	0.80
	Seguridad	0.97	0.96	0.95
	Servidores	0.68	0.58	0.60
	Sistema Operativo	0.41	0.37	0.34
	Software	0.72	0.72	0.70
	Tecnolex	0.64	0.61	0.57
	Tratamiento de Archivos	0.82	0.84	0.77
	Videollamada	0.00	0.17	0.12
	none	0.36	0.26	0.15

Como se puede apreciar en la tabla, se cumple lo indicado previamente.

8.5.2. Segundo dataset

Mientras tanto, para el segundo dataset creado sus resultados se pueden revisar en el Anexo E de este informe. Como se puede apreciar, a pesar de que se ha decidido crear 2 clases nuevas para este dataset, los modelos tienen a presentar buenas métricas para las clases, a excepción de la clase Consulta, la cual como se puede observar en las matrices de confusión, se ve afectada por la clase Requerimiento, lo cual se puede explicar por el desbalance que existe en el dataset. A la vez, las matrices para esta data set se encuentran presentes en el Anexo F de este informe.

Se presenta la siguiente tabla en donde se puede ver la comparación entre las métricas obtenidas al correr en las distintas bases.

Base	Clase	Métrica	Regresión	SVM	Árbol
1ra base de datos	Consulta	Precision	0.55	0.52	0.40
		Recall	0.26	0.29	0.31
		F1-Score	0.36	0.37	0.35
		Support	2055	2055	2055
	Incidente	Precision	0.64	0.67	0.48
		Recall	0.32	0.29	0.40
		F1-Score	0.43	0.40	0.44
		Support	1610	1610	1610
	Requerimiento	Precision	0.88	0.88	0.89
		Recall	0.97	0.96	0.92
		F1-Score	0.92	0.92	0.91
		Support	19702	19702	19702
2da base de datos	Consulta	Precision	0.73	0.64	0.68
		Recall	0.48	0.58	0.50
		F1-Score	0.58	0.61	0.58
		Support	165	165	165
	Incidente	Precision	0.80	0.78	0.71
		Recall	0.76	0.75	0.66
		F1-Score	0.78	0.77	0.69
		Support	420	420	420
	Requerimiento	Precision	0.83	0.86	0.80
		Recall	0.90	0.87	0.86
		F1-Score	0.86	0.86	0.83
		Support	1123	1123	1123

En la tabla se puede apreciar como las métricas para las clases Consulta e Incidente aumentan al contar con más tickets, pero para la clase Requerimiento este valor disminuye al haber cambiado tickets asignados con este valor a otras opciones.

8.5.3. Tercer dataset

A la vez, si se revisan los resultados al trabajar los modelos para el atributo Categoría del tercer data set creado, los cuales se encuentran presentes en el Anexo G, podemos observar que, al haber eliminado clases, asignando sus tickets a otras clases existentes o a la nueva clase creada, No categorizable, los modelos tienden a presentar buenos resultados a la hora de buscar categorizar los tickets, siendo las clases con peores resultados las que cuentan con un menor número de tickets designados. Esto se puede apreciar en las matrices de confusión presentes en el Anexo H. De los resultados obtenidos al trabajar con los dataset 2 y 3, se ha podido descubrir los siguiente. El eliminar o agregar clases puede tener un gran efecto sobre los modelos a la hora de querer clasificar los tickets. También se puede apreciar que los modelos tienden a asignar a los tickets las clases con el mayor número de valores, por lo que la idea de aplicar técnicas de balanceo sirve a la hora de buscar reducir el efecto previamente mencionado.

8.5.4. Cuarto dataset

Esto se puede apreciar en mayor detalle en los resultados obtenidos al aplicar los modelos para la cuarta y quinta base de datos, las cuales han pasado por un proceso de balanceo de los datos. Todos estos resultados se encuentran disponibles en los Anexos I-L. Si estudiamos los resultados obtenidos para los modelos al correrlos sobre la cuarta base, podemos observar que haber balanceado al eliminar datos de forma aleatoria hasta el punto que todas las cuentan con la misma cantidad de tickets permite mejorar la clasificación de los mismos por parte de los modelos. Inclusive en este caso ocurre que ahora la clase que presenta mejores resultados corresponde a Incidente y no Requerimiento, lo que a entender que se les hace más fácil a los modelos reconocer que tickets corresponden a esa clase comparada con los otras dos.
Aun así, se debe hacer notar que este caso no tiene sentido a la hora de aplicarlo a la situación real, pues se ha podido observar que inclusive creando nuevas clases el desbalance para el atributo Tipo Requerimiento sigue estando presente, en donde la clase Requerimiento presenta una mayor cantidad de tickets comparada con el resto de clases. Se desea notar que este trabajo de balanceo fue creado con el fin de estudiar la capacidad de los modelos para clasificar con los tickets si se contara con una base de datos balanceada, y como se ha podido apreciar presentan mejores resultados que en el caso opuesto.

8.5.5. Quinto dataset

Mientras tanto, para la quinta base de datos creada también se aplicó una técnica para balancear las clases, la cual consistió en crear nuevas clases más generales que permitieran agrupar las ya existentes, en donde se pasó de un atributo con 14 clases a 5. De esta forma, se buscó crear nuevas clases que contaran con una mayor cantidad de tickets y de esta forma equiparar el número de datos entre las clases. Como se puede observar los modelos presentan mucho mejores resultados para este atributo comparado con los resultados obtenidos por los modelos en el primer data set, en donde se intentó clasificar usando todas las clases originales. Se debe notar que para la base balanceada los modelos tienden a asignar tickets como Software en casos que no corresponde, lo cual se puede explicar porque esta sigue siendo la clase del atributo Categoría con el mayor número de tickets para todos los data set creados. A pesar de que esta situación posee sentido en la aplicación real, al ser Tecnolex una empresa que ofrece servicios informáticos como desarrollo y mantención de Software, sería interesante aplicar una nueva técnica de balanceo que permita reducir aún más el número de tickets que se asigna como Software, con el fin de disminuir aún más el desbalance en los datos.

Con respecto a los modelos implementados, se puede apreciar que el modelo que presenta los mejores resultados a través de todos los data set corresponde a la Regresión Logística, lo cual presenta interesante observaciones acerca de la naturaleza de los datos y la estructura de los data set creados, ya que uno esperaría que al trabajar con atributos multiclase el modelo SVM sería el que presente mejores resultados. Estos resultados también dan a entender que la relación entre las clases y el texto de los tickets es una relación lineal, lo cual tiene sentido ya que dependiendo

9. Conclusiones

Para este trabajo se buscaba obtener un modelo de clasificación que permitiera entregar una sugerencia a los técnicos de Helpdesk respecto a las clasificaciones a los tickets para los atributos Tipo Requerimiento y Categoría. Durante la realización de ese objetivo se lograron obtener observaciones respecto al trato de los técnicos respecto al registro de información del área y la capacidad de un modelo de machine learning en impactar en la clasificación de textos. En el siguiente listado se puede apreciar las principales conclusiones a las que se llegó de los resultados obtenidos durante la realización de este trabajo.

9.1. Efecto de la calidad de los datos: Durante el proceso de limpieza y transformación de los datos, se encontraron tickets que contaban con muy poca información o completamente vacíos, se encontraron tickets con origen de correo los cuales poseían información que no servía a la hora de querer clasificarlos, como las firmas de las empresas o los nombres y correos de las personas que levantaron el ticket, se encontraron tickets que no poseían ninguna clasificación de cualquier tipo, para tickets con origen por teléfono se encontraron varios en donde la única información que se registraba era el número de teléfono de la persona que levanto la atención, entre otros casos. Todas estas situaciones provocaron un extra de trabajo a la hora de querer correr los modelos, pues provocaban ruido en los resultados. Se da a entender que se debería tener cuidado a la hora de ingresar la información a los tickets, pues esta puede llegar a tener un gran impacto a la hora de querer clasificar los tickets.

9.2. Bases desbalanceadas sobre los modelos: Como se logró apreciar en los resultados obtenidos al correr los modelos para todas las bases de datos creadas, el desbalanceo de los datos llega a tener un gran impacto a la hora de querer clasificar los tickets, en donde los modelos tienden a asignarle a los tickets la clase con el mayor número de valores. Por lo tanto, a la hora de querer trabajar con las bases podría ser una buena idea crear nuevas clases que permitieran agrupar de mejor forma los tickets e intentar bajar el grado de desbalanceo en los datos.

9.3. Distintos modelos de Machine Learning: Con respecto a la hora de elegir los modelos a entrenar, se debe tomar en cuenta la naturaleza de los datos y la estructura de la base de datos con la que se trabaje, como el nivel de desbalance de la misma y el número de clases. Con respecto a los modelos utilizados, se puede apreciar que el modelo que entrega las mejores métricas corresponde al modelo de Regresión Logística, lo que es un resultado interesante, ya que uno esperaría que los modelo SVM y árbol de Clasificación presentarían mejores resultados al trabajar con atributos de múltiples clases, desbalanceadas y con una base con gran cantidad de data.

9.4. Hipótesis planteadas: Con respecto a las hipótesis planteadas, no se lograría afirmar que no existe una capacitación suficiente de los técnicos, pero si se podría llegar a la conclusión que no se le dedica el tiempo suficiente, ya que se ha podido apreciar que las distribuciones de tickets se mantienen bastante similares a través del tiempo y se han podido encontrar tickets a los cuales no se les ha ingresado información es bastante pobre, como solo el número de teléfono de la persona que solicita la atención. Esto da a entender que no se tiene el cuidado suficiente por parte de los técnicos a la hora de escribir la información en los atributos respectivos de los tickets. Con respecto a una falta de estructura por parte de Tecnolex para poder cuidar el ingreso de información a los tickets, como se ha visto previamente el atributo de los tickets con la mayor cantidad de valores nulos corresponde a Tiempo_Respuesta, el cual es un valor que debería ser ingresado por el sistema SysAid de forma automática, lo que da a entender que hay un problema tecnológico ahí presente que se debería arreglar.

9.5. Definición de nuevas clases para los atributos: Como se ha visto durante el trabajo, el desbalance de los datos puede afectar de forma negativa a los modelos a la hora de querer clasificar los tickets, y una de las técnicas que se utilizó que presento buenos resultados fue crear nuevas clases para clasificar los tickets, por lo que sería buena idea por parte de Tecnolex definir nuevas clases con el fin de eliminar otras que se encuentren obsoletas y puedan ser englobadas con una nueva, y de paso reducir el número de clases y balancear la base, facilitando el trabajo de clasificar los tickets.

9.6. Veracidad de la declaración realizada por el jefe de área: Uno de los factores principales para comenzar este trabajo fue la declaración de unos de los jefes del área, quien indicada que las clasificaciones que se les habían designado a los tickets más recientes no eran las correctas, por lo que no se podían realizar los reportes solicitados por la Alta Gerencia. Se logró descubrir que un gran porcentaje de los tickets estaban realmente mal clasificados, por lo que se recomendaría hacer una revisión manual de los tickets, con el fin de contar con información de no confianza, pues uno a priori no sabría que tickets están bien clasificados.

9.7. Aplicación de los modelos a la hora de sugerir clasificaciones para los atributos: Como se ha podido apreciar durante el trabajo, los modelos de machine learning son herramientas que sirven a la hora de querer clasificar textos en distintas clases, pero estos se ven afectados por la naturaleza de los datos, su desbalance y el nivel de clases que deben ser clasificadas, por lo tanto, se debería tomar en cuenta estos puntos a la hora de trabajar con modelos de machine learning. Específicamente para este problema se debería tomar en gran cuenta el desbalance para el atributo Tipo Requerimiento, ya que, al ser originalmente una variable con 3 clases la estrategia de eliminar clases con pocos datos al agruparlas en clases más generales no tiene sentido, pero se podría probar una estrategia inversa, que consistiría en crear nuevas clases que permitan clasificar ticket con valor Requerimiento, pero que los técnicos no estén tan seguros, en nuevas clases mejor representativas, y de esta forma balanceando la base de datos.

10. Recomendaciones y trabajos futuros

Por lo tanto, sumando todo lo visto en este trabajo, se presenta el siguiente listado de recomendaciones para Tecnolex y posibles trabajos futuros que se podrían realizar para mejor el hecho durante esta memoria.

10.1. Realizar una nueva revisión de la base de datos: Como se logró apreciar en los dataset en los cuales se realizó una revisión manual, existan tickets que se encuentran en una clase incorrecta, lo que termina por dificultar a los modelos a la hora de querer clasificar los tickets. Esto sería un trabajo de largo tiempo, ya que el trabajo de revisar los sub sets creados tomo un tiempo de 5 semanas, pero sería bastante provechoso para el área, pues les permitiría contar con información de confianza.

10.2. Revisión del sistema SysAid: Se ha podido notar que ciertos atributos de los tickets que deberían ser ingresados de forma automática por parte del sistema cuentan con valores nulos, por lo que sería buena idea realizar una revisión de cómo está llenando los campos de los tickets SysAid.

10.3. Implementación de los modelos: Se considera que sería una buena idea utilizar los modelos como una herramienta de apoyo y sugerencia para los técnicos, pero sería también provechoso implementar más control a los técnicos en su forma de llenar los atributos de los tickets.

10.4. Desarrollo de un nuevo modelo: Se considera que ha futuro se podría estudiar implementar un modelo de redes neuronales, debido a la naturaleza de la base de datos, y compararlo con el modelo de Regresión Logística, con el fin de concluir completamente respecto a la complejidad del dataset. Con respecto a porque no se hizo uso de una red neuronal durante este trabajo, se debe al largo tiempo que toma implementar una red neuronal y por la situación particular de este trabajo.

11. Bibliografía

Cáceres, P. A. (2020). *DISEÑO Y CONSTRUCCIÓN DE MODELO DE CLASIFICACIÓN DE INCIDENTES DE SEGURIDAD USANDO NLP EN LOS REGISTROS DE TEXTO ESCRITO PARA AUTOMARIZAR ETIQUETACIÓN.* Santiago de Chile.

Fayyad, U. P.-S. (1996). *From data mining to knowledge discovery in databases.*

IBM. (03 de 01 de 2023). *IBM*. Obtenido de https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-classification-tree

Innnovation, A. (22 de Octubre de 2019). *atriainnovation*. Obtenido de https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

KeepCoding, R. (8 de febrero de 2023). *keepcoding.io*. Obtenido de https://keepcoding.io/blog/que-es-la-lematizacion-en-python/#:~:text=La%20lematizaci%C3%B3n%20en%20Python%20es,%27%20%E2%86%92%20%27entrenar%27).

mathworks. (s.f.). *la.mathworks.com*. Obtenido de https://la.mathworks.com/discovery/support-vector-machine.html

Redondo, Jose Antonio Martin. (s.f.). *blogs.imf-formacion.* Obtenido de https://blogs.imf-formacion.com/blog/tecnologia/glosario-de-procesamiento-de-lenguaje-natural-nlp-202302/#:~:text=Procesamiento%20de%20Lenguaje%20Natural%20(NLP)%3A%20Es%20una%20rama%20de,como%20palabras%2C%20frases%20o%20oraciones.

***Unioviedo.* (s.f.). Obtenido de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html#:~:text=K%2Dmeans%20es%20un%20algoritmo,de%20su%20grupo%20o%20cluster.**

Votti, E. (09 de Marzo de 2021). *medium*. Obtenido de https://medium.com/idatha-enterprise-experience-academic-s/hablemos-de-zero-shot-learning-db52bd558e73

12. Anexos

Anexo A: Classification Report para atributo Tipo de Requerimiento de la primera base de datos

		Regresión	SVM	Árbol Decisión
Consulta	Precision	0.55	0.52	0.40
	Recall	0.26	0.29	0.31
	F1-Score	0.36	0.37	0.35
	Support	2055	2055	2055
Incidente	Precision	0.64	0.67	0.48
	Recall	0.32	0.29	0.40
	F1-Score	0.43	0.40	0.44
	Support	1610	1610	1610
Requerimiento	Precision	0.88	0.88	0.89
	Recall	0.97	0.96	0.92
	F1-Score	0.92	0.92	0.91
	Support	19702	19702	19702

Tabla 3: Classification Report para los modelos aplicados a Tipo Requerimiento

Anexo B: Matrices de confusión normalizada para atributo Tipo Requerimiento de la primera base de datos

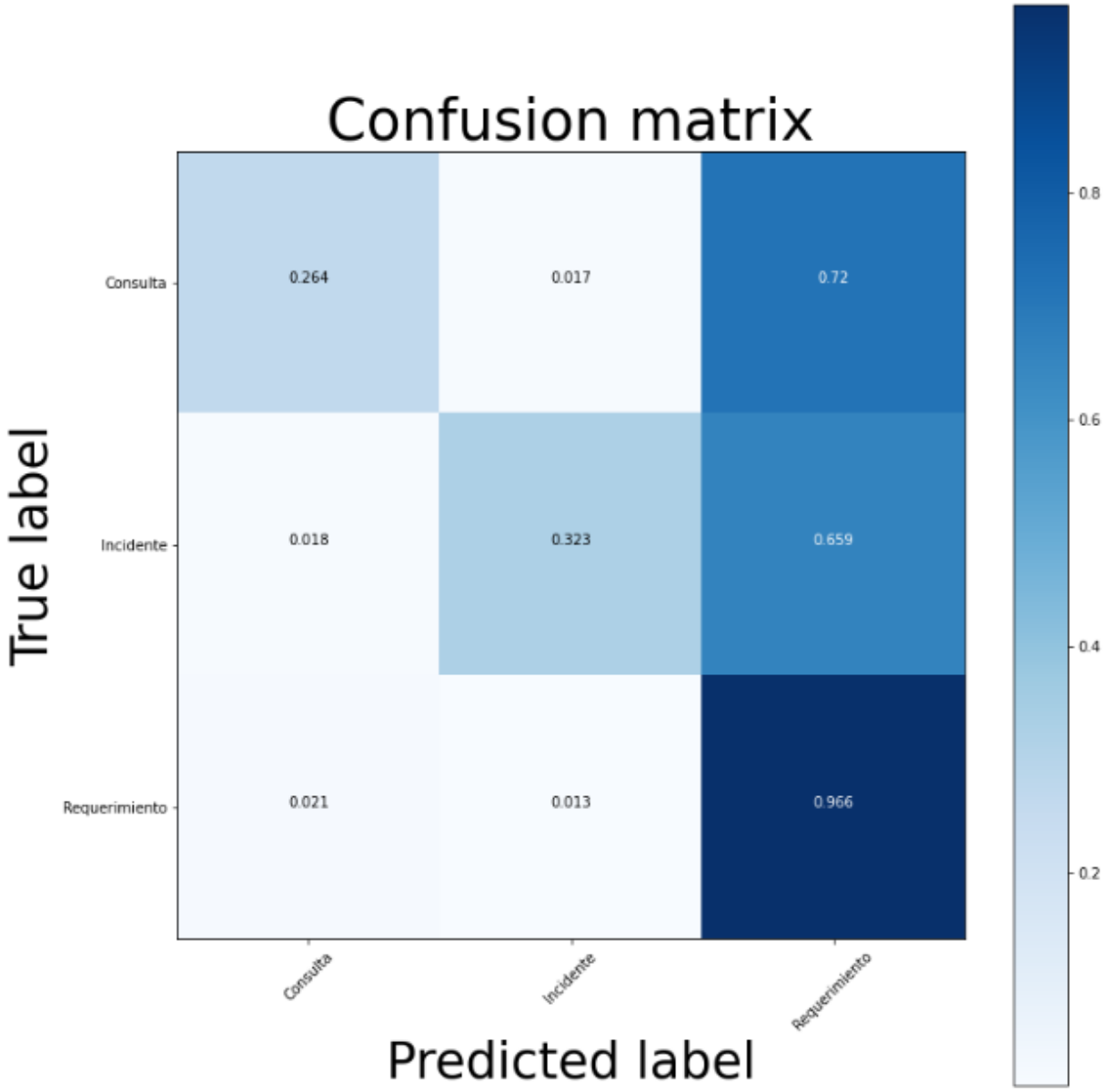

Figura 20: Matriz de confusión normalizada para la Regresión Logística

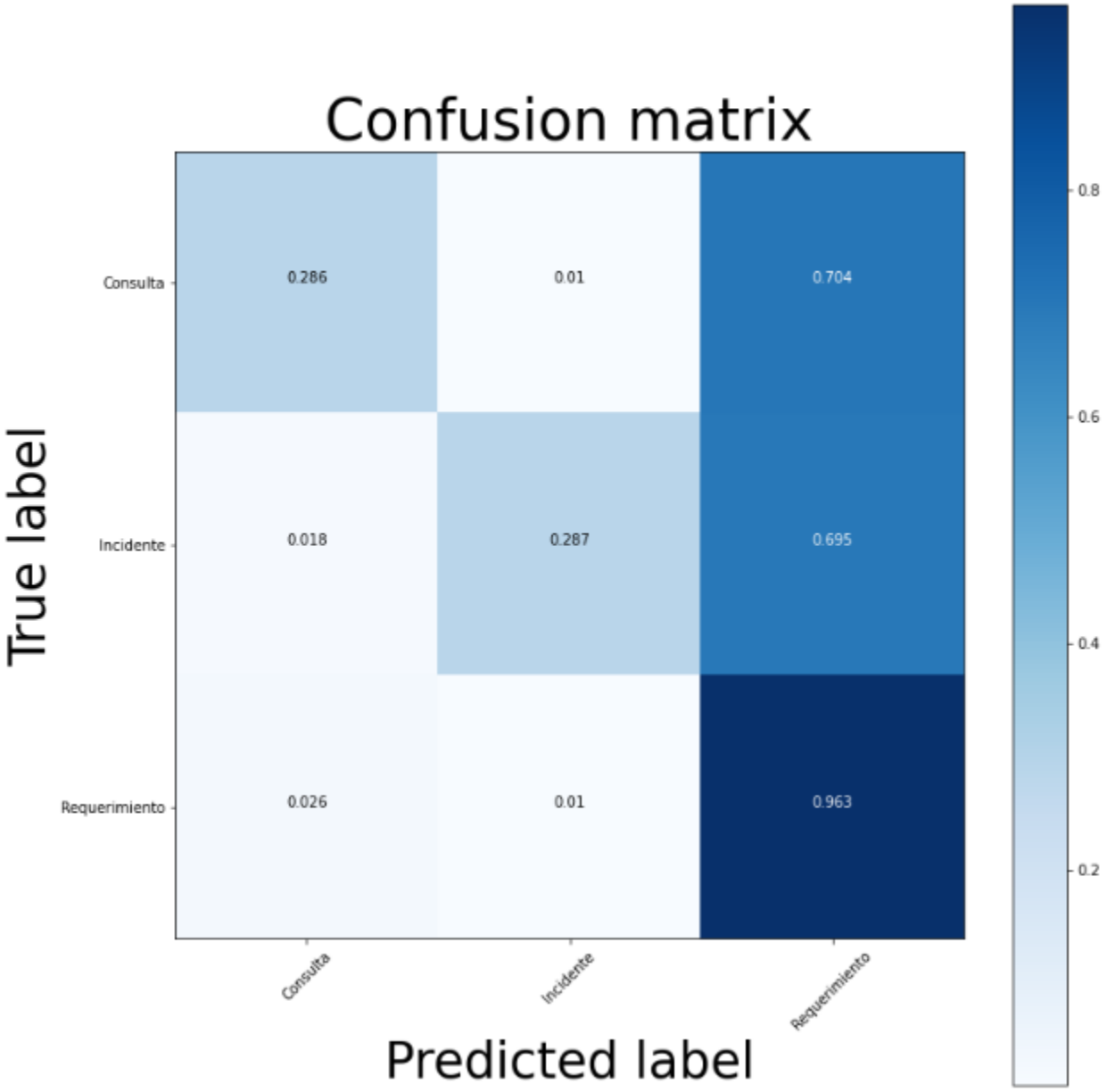

Figura 21: Matriz de confusión normalizada para el SVM

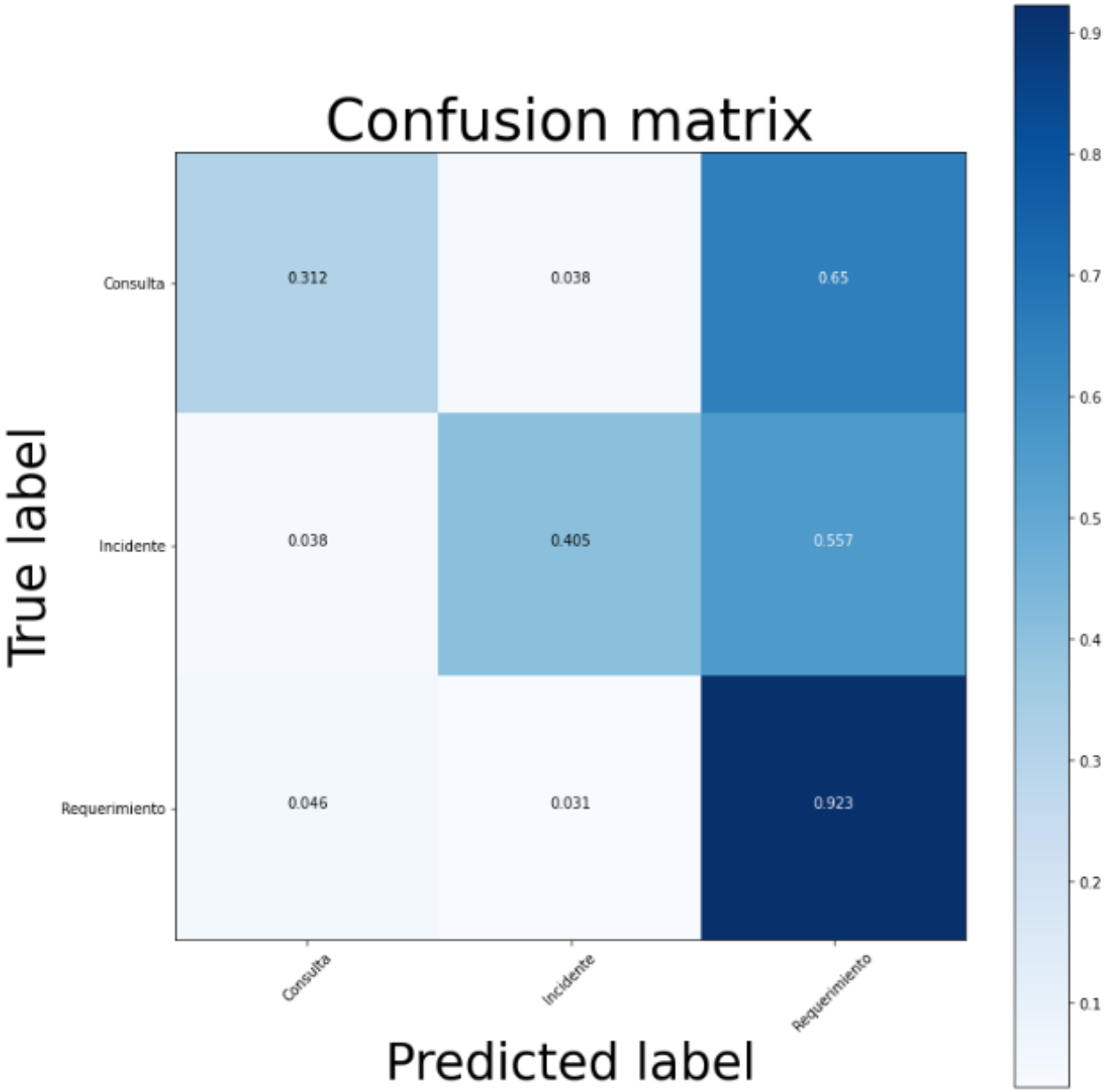

Figura 22: Matriz de confusión normalizada para el Árbol de clasificación

Anexo C: Classification Report para atributo Categoría de la primera base de datos

		Regresión	SVM	Árbol Decisión
Acceso	Precision	0.73	0.69	0.65
	Recall	0.65	0.65	0.62
	F1-Score	0.69	0.67	0.64
	Support	2244	2244	2244
Comunicaciones	Precision	0.00	0.00	0.00
	Recall	0.00	0.00	0.00
	F1-Score	0.00	0.00	0.00
	Support	3	3	3
Hardware	Precision	0.72	0.70	0.65
	Recall	0.69	0.68	0.63
	F1-Score	0.70	0.69	0.64
	Support	3037	3037	3037
Por Definir	Precision	0.30	0.26	0.25
	Recall	0.12	0.16	0.14
	F1-Score	0.17	0.20	0.18
	Support	50	50	50
Red	Precision	0.74	0.72	0.64
	Recall	0.55	0.57	0.53
	F1-Score	0.63	0.63	0.58
	Support	583	583	583
Sala de Reunión	Precision	0.87	0.86	0.80
	Recall	0.85	0.87	0.82
	F1-Score	0.86	0.86	0.81
	Support	1104	1104	1104
Seguridad	Precision	0.97	0.96	0.95
	Recall	0.98	0.98	0.96
	F1-Score	0.99	0.97	0.95
	Support	313	313	313
Servidores	Precision	0.68	0.58	0.60
	Recall	0.58	0.60	0.50
	F1-Score	0.62	0.59	0.55
	Support	180	180	180
Sistema Operativo	Precision	0.41	0.37	0.34
	Recall	0.18	0.17	0.21
	F1-Score	0.25	0.23	0.23
	Support	619	619	619
Software	Precision	0.72	0.72	0.70
	Recall	0.84	0.82	0.76
	F1-Score	0.78	0.77	0.73
	Support	9008	9008	9008

Tecnolex	Precision	0.64	0.61	0.57
	Recall	0.41	0.42	0.42
	F1-Score	0.50	0.50	0.48
	Support	515	515	515
Tratamiento de Archivos	Precision	0.82	0.84	0.77
	Recall	0.83	0.80	0.78
	F1-Score	0.82	0.82	0.77
	Support	5386	5386	5386
Videollamada	Precision	0.00	0.17	0.12
	Recall	0.00	0.03	0.06
	F1-Score	0.00	0.05	0.08
	Support	32	32	0.32
none	Precision	0.36	0.26	0.15
	Recall	0.10	0.09	0.08
	F1-Score	0.16	0.14	0.11
	Support	275	275	275

Tabla 4: Classification Report para los modelos aplicados a Categoría

Anexo D: Matrices de confusión normalizada para atributo Categoría de la segunda base de datos

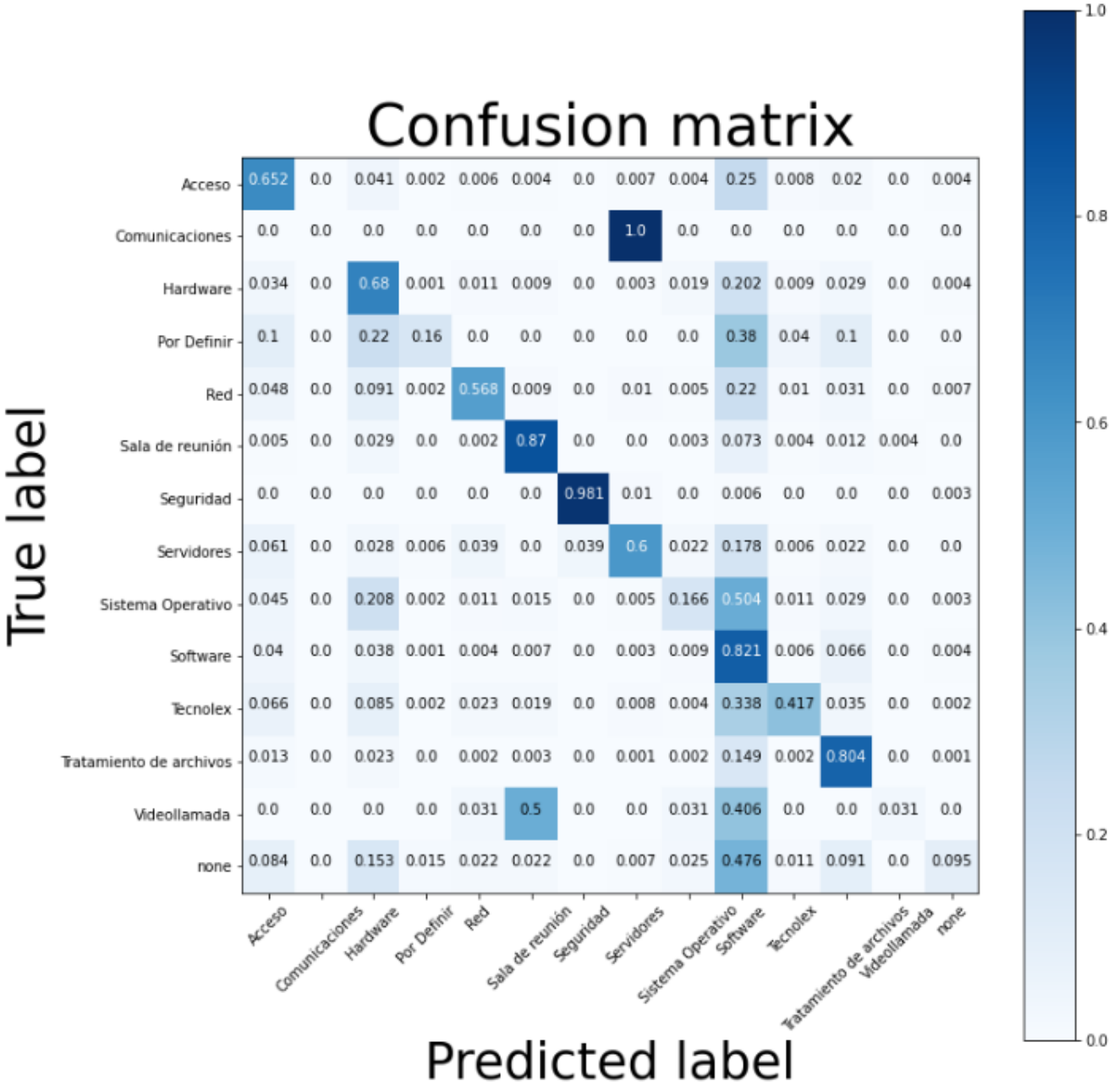

Figura 23: Matriz de confusión normalizada para la Regresión Logística

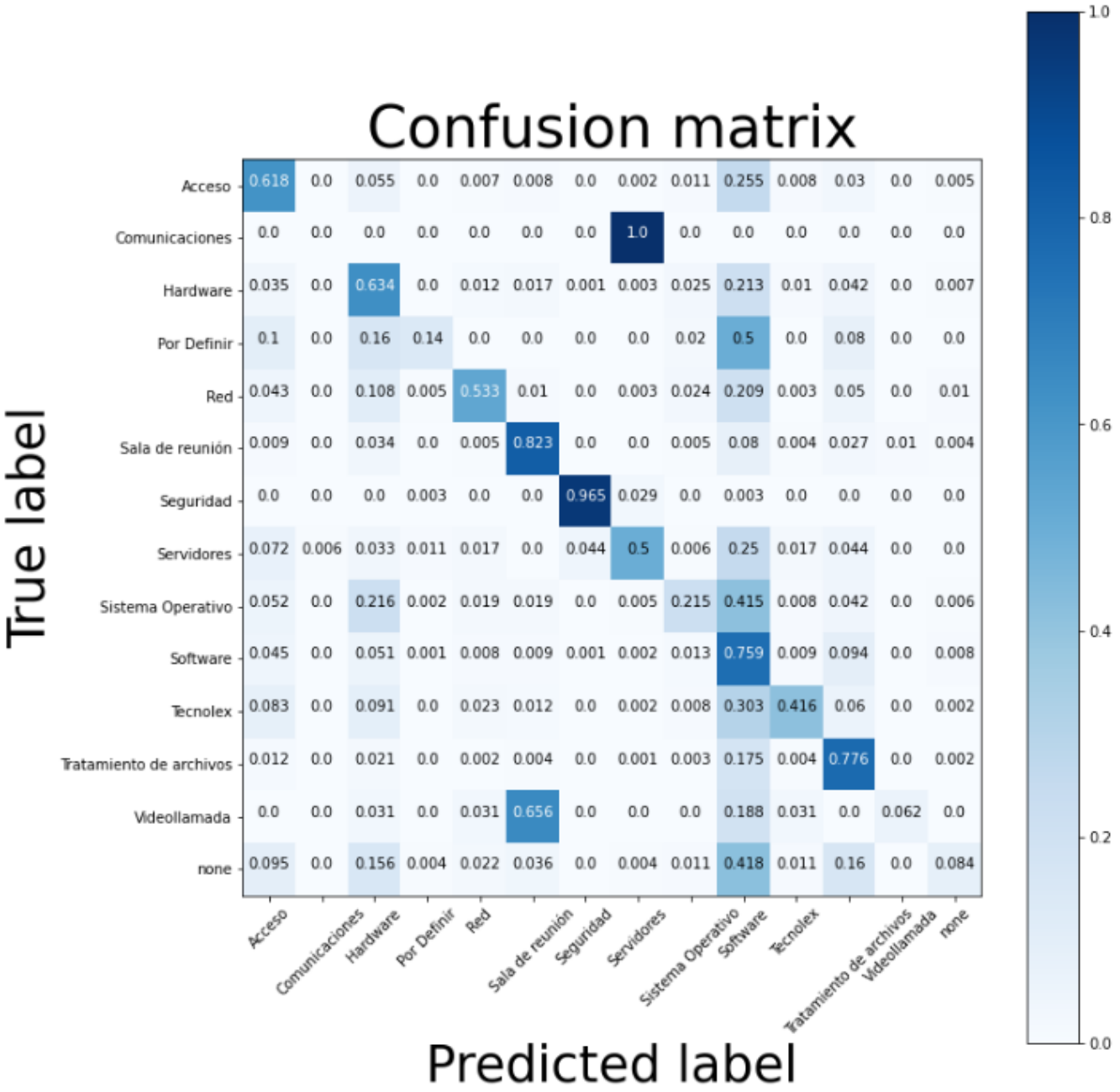

Figura 24: Matriz de confusión normalizada para el SVM

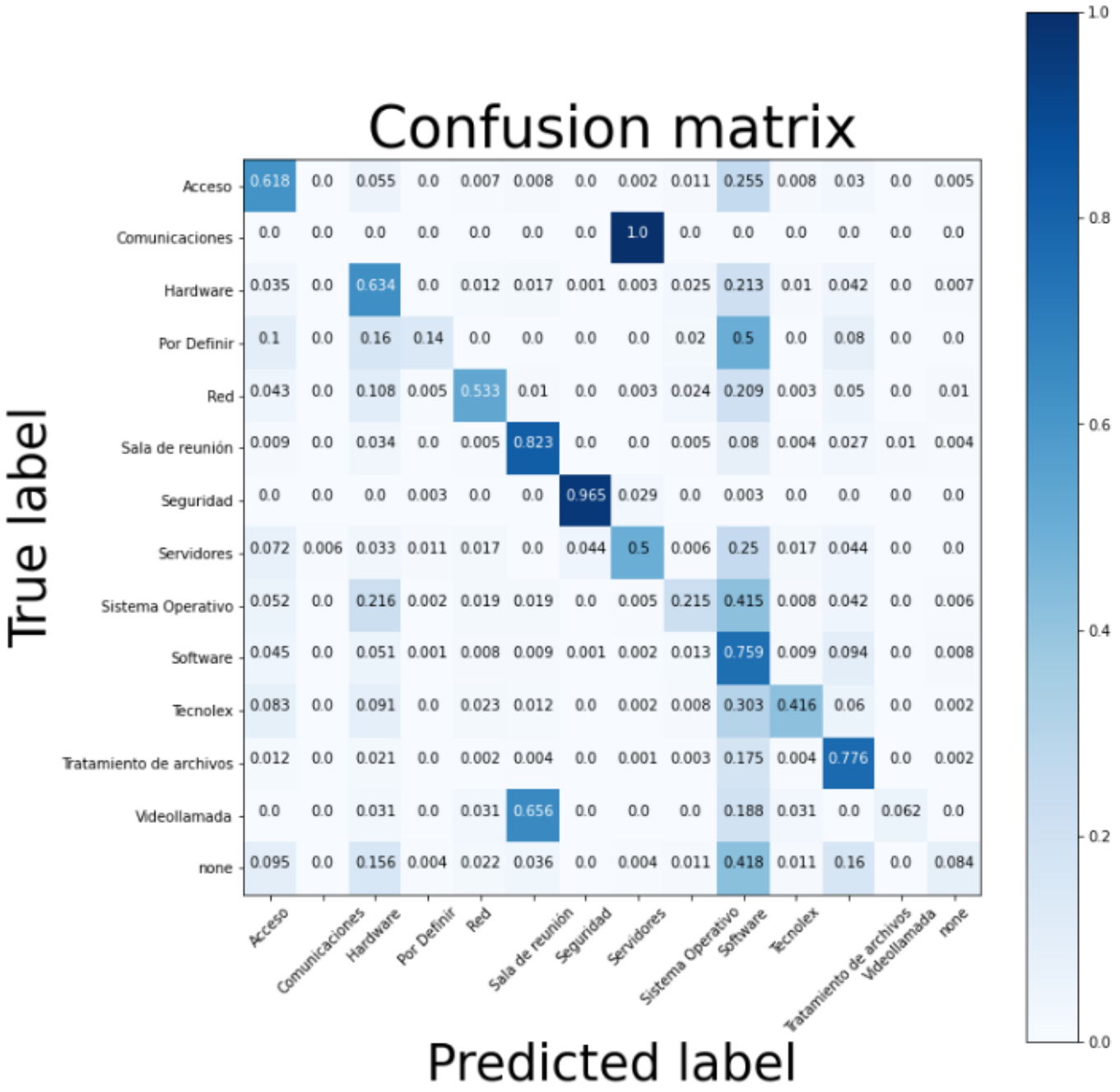

Figura 25: Matriz de confusión normalizada para el Árbol de clasificación

Anexo E: Classification Report para atributo Tipo Requerimiento de la segunda base de datos

		Regresión	SVM	Árbol Decisión
Consulta	Precision	0.73	0.64	0.68
	Recall	0.48	0.58	0.50
	F1-Score	0.58	0.61	0.58
	Support	165	165	165
Incidente	Precision	0.80	0.78	0.71
	Recall	0.76	0.75	0.66
	F1-Score	0.78	0.77	0.69
	Support	420	420	420
Informativo	Precision	0.83	0.74	0.70
	Recall	0.62	0.67	0.60
	F1-Score	0.71	0.71	0.59
	Support	146	146	146
No Categorizable	Precision	0.60	0.60	0.57
	Recall	0.69	0.71	0.60
	F1-Score	0.65	0.65	0.59
	Support	146	146	146
Requerimiento	Precision	0.83	0.86	0.80
	Recall	0.90	0.87	0.86
	F1-Score	0.86	0.86	0.83
	Support	1123	1123	1123

Tabla 5: Classification Report para los modelos aplicados a Tipo Requerimiento

Anexo F: Matrices de confusión normalizada para atributo Tipo Requerimiento de la segunda base de datos

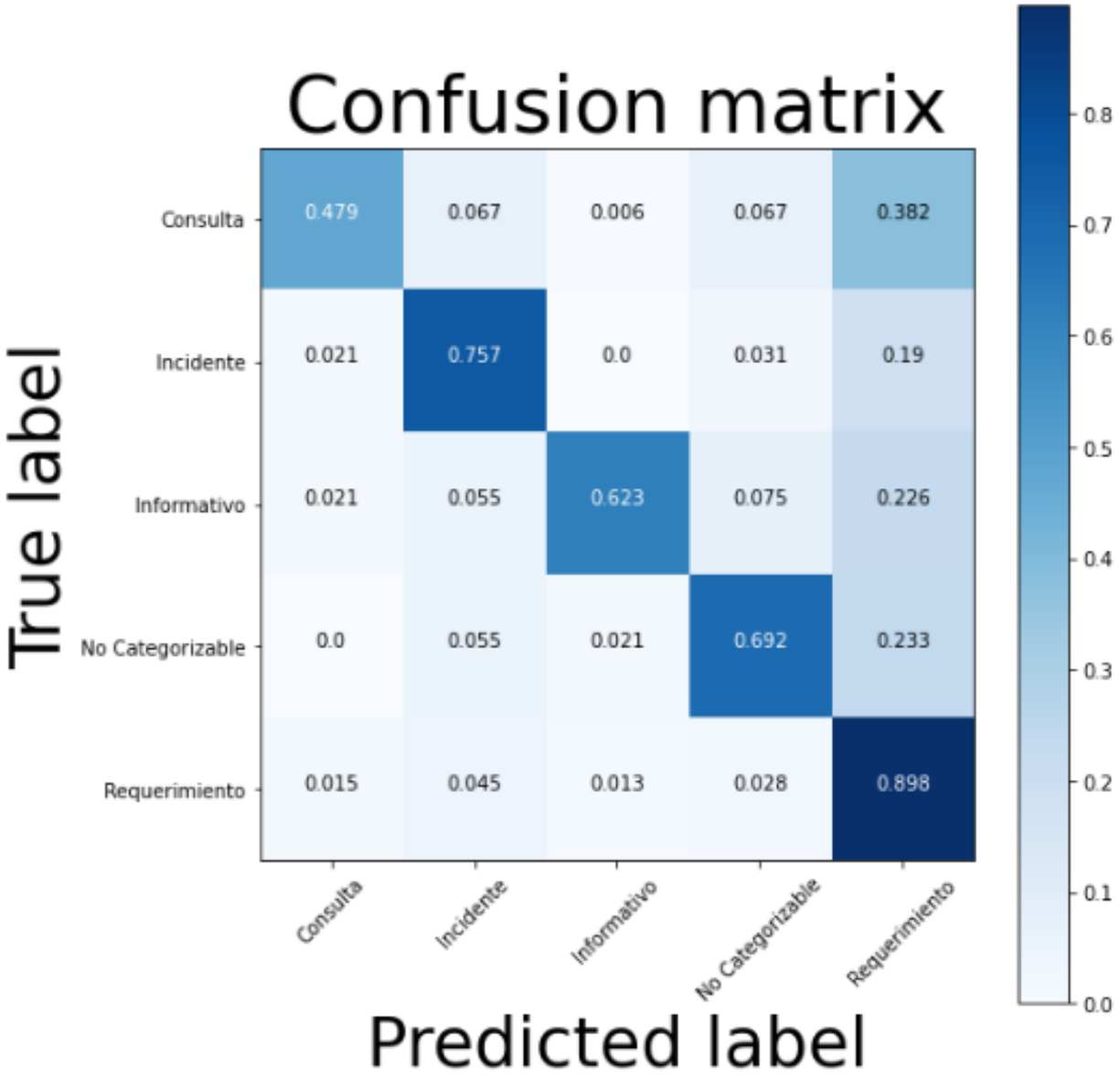

Figura 26: Matriz de confusión normalizada para la Regresión Logística

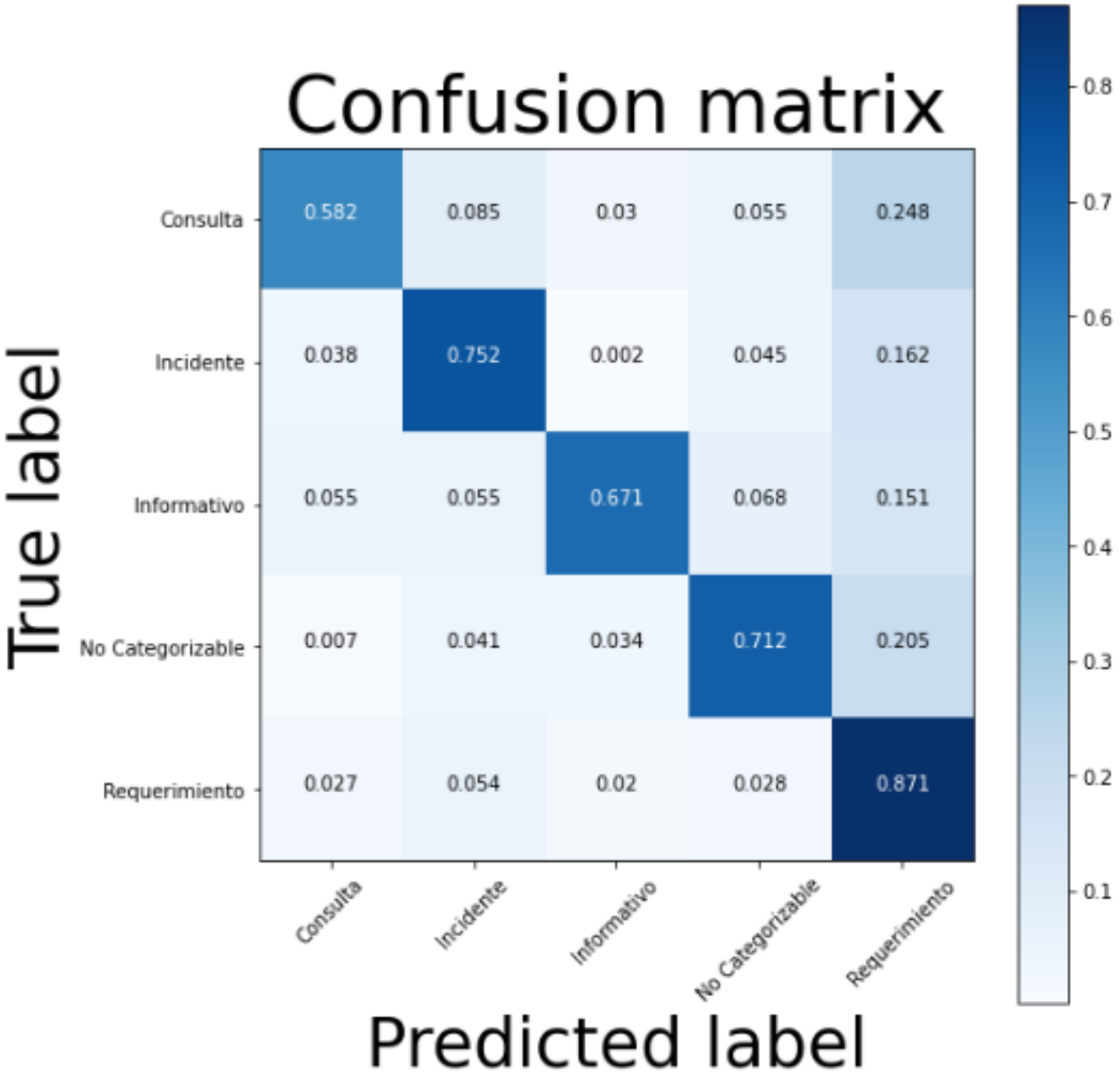

Figura 27: Matriz de confusión normalizada para el SVM

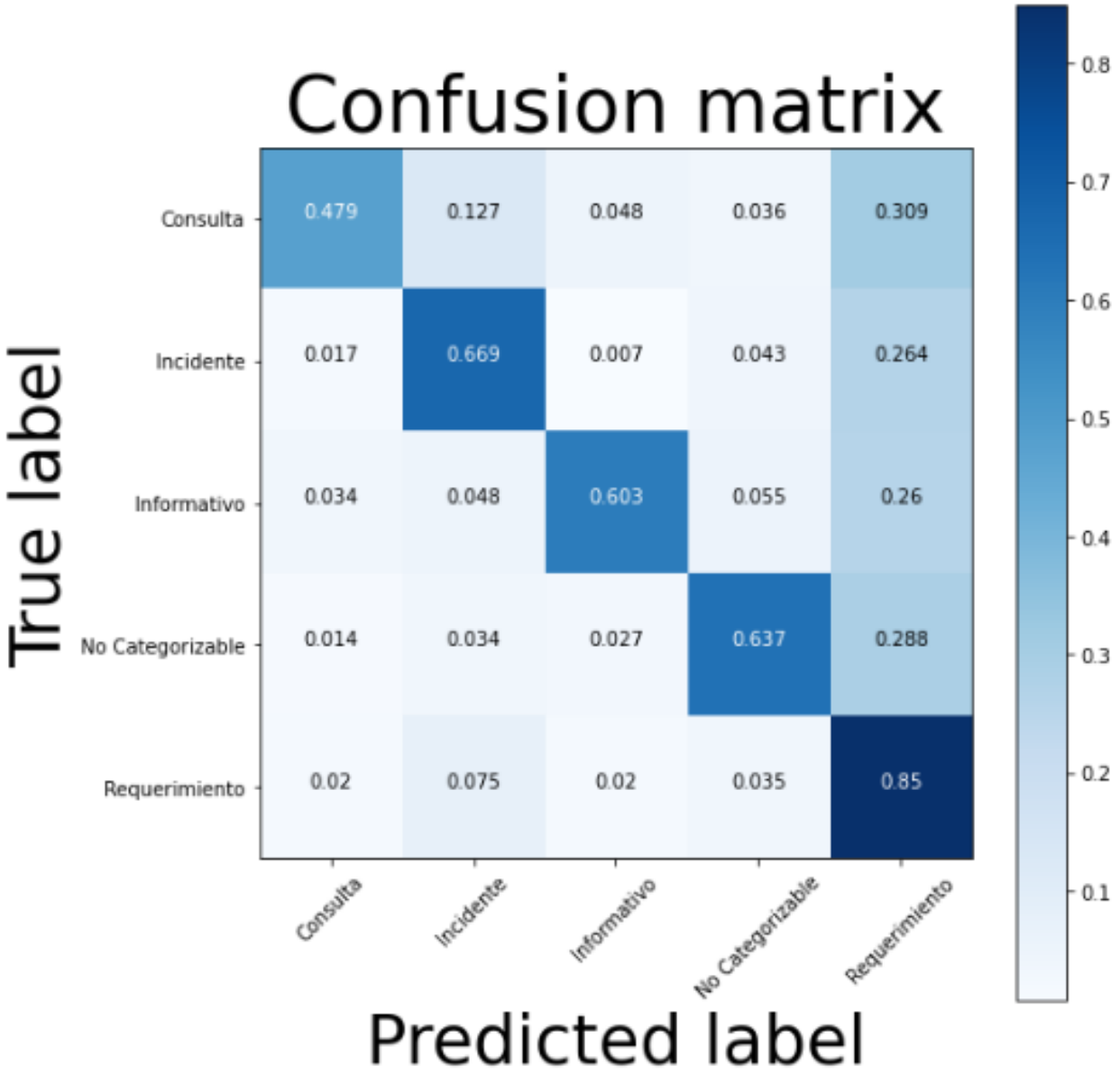

Figura 28: Matriz de confusión normalizada para el Árbol de clasificación

Anexo G: Classification Report para atributo Categoría de la tercera base de datos

		Regresión	SVM	Árbol Decisión
Acceso	Precision	0.74	0.68	0.74
	Recall	0.65	0.63	0.63
	F1-Score	0.69	0.65	0.68
	Support	82	82	82
Hardware	Precision	0.71	0.70	0.49
	Recall	0.59	0.62	0.48
	F1-Score	0.65	0.66	0.49
	Support	69	69	420
No Categorizable	Precision	0.40	0.31	0.28
	Recall	0.34	0.49	0.57
	F1-Score	0.37	0.38	0.38
	Support	35	35	35
Red	Precision	1.00	0.89	0.81
	Recall	0.59	0.73	0.59
	F1-Score	0.74	0.80	0.68
	Support	22	22	22
Sala de Reunión	Precision	0.92	0.92	0.86
	Recall	0.79	0.82	0.86
	F1-Score	0.85	0.87	0.86
	Support	28	28	28
Seguridad	Precision	1.00	1.00	1.00
	Recall	0.85	0.85	0.85
	F1-Score	0.92	0.92	0.92
	Support	13	13	13
Servidores	Precision	1.00	0.60	0.50
	Recall	0.33	1.00	0.33
	F1-Score	0.50	0.75	0.40
	Support	3	3	3
Software	Precision	0.67	0.71	0.58
	Recall	0.85	0.74	0.58
	F1-Score	0.75	0.72	0.58
	Support	186	186	186
Tecnolex	Precision	0.68	0.62	0.35
	Recall	0.45	0.45	0.18
	F1-Score	0.55	0.53	0.24
	Support	33	33	33
Tratamiento de archivos	Precision	0.82	0.83	0.76
	Recall	0.85	0.83	0.78
	F1-Score	0.84	0.83	0.77
	Support	129	129	129

Tabla 6: Classification Report para los modelos aplicados a Categoría

Anexo H: Matrices de confusión normalizada para atributo Categoría de la tercera base de datos

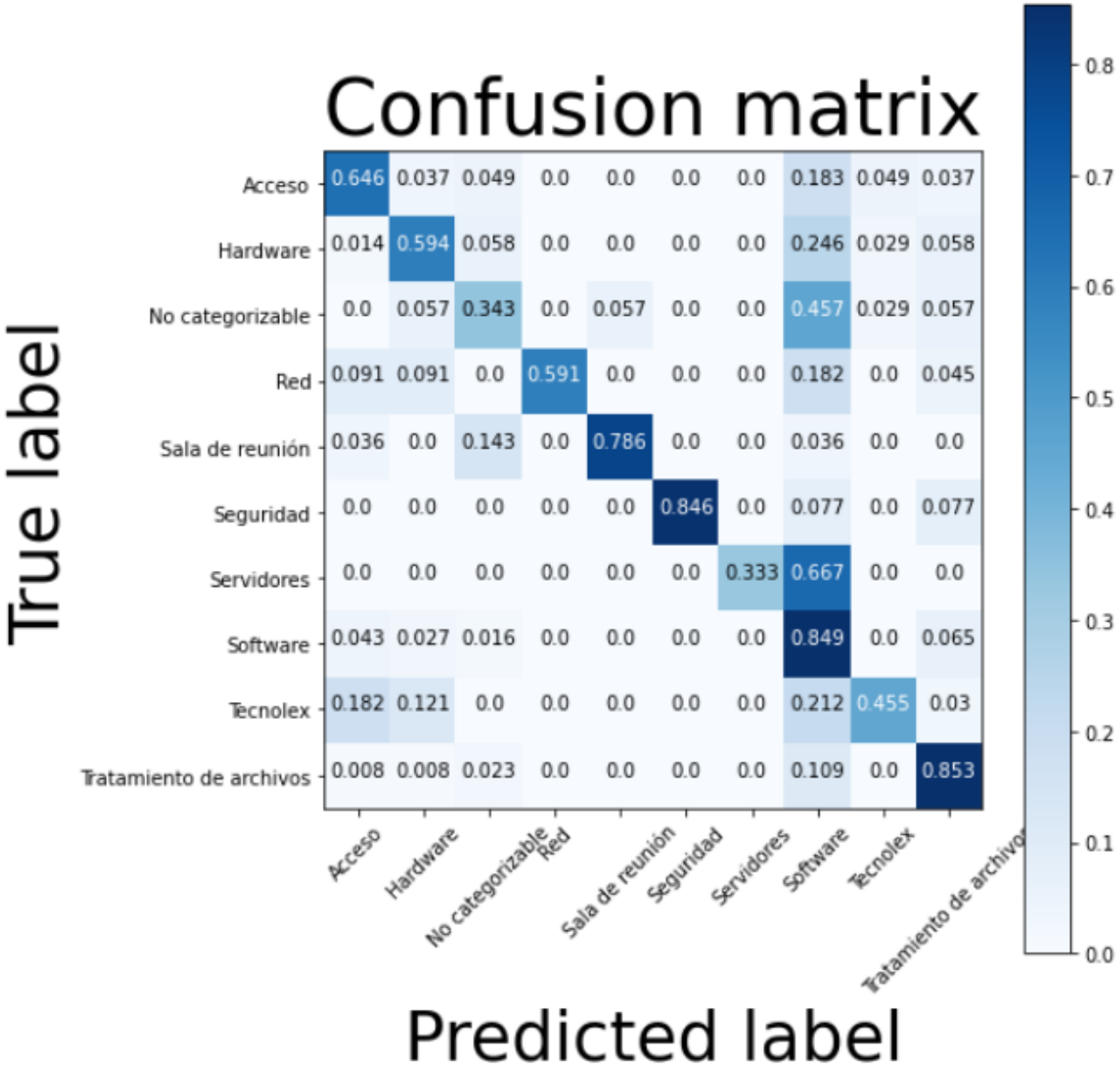

Figura 29: Matriz de confusión normalizada para la Regresión Logística

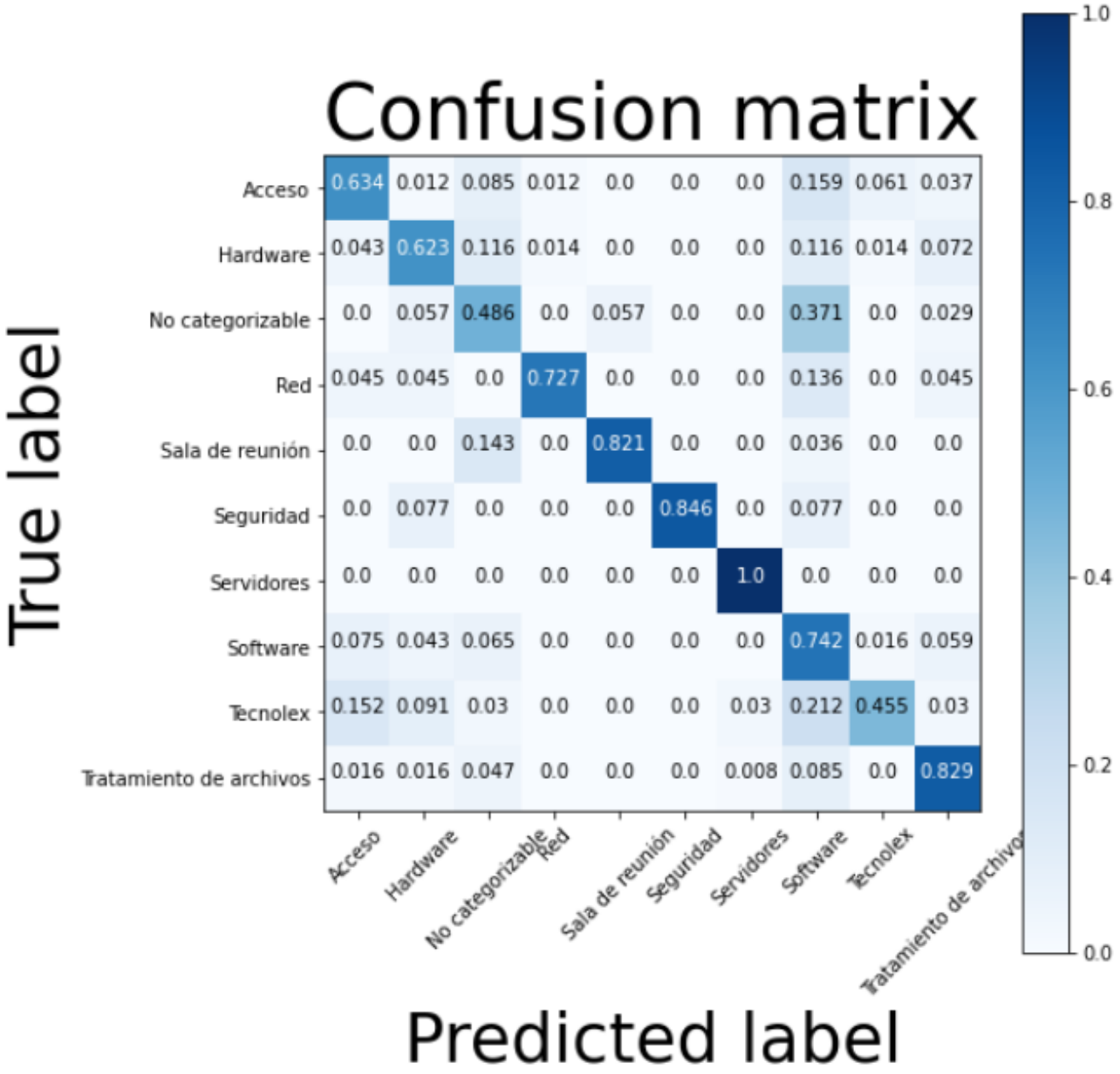

Figura 30: Matriz de confusión normalizada para el SVM

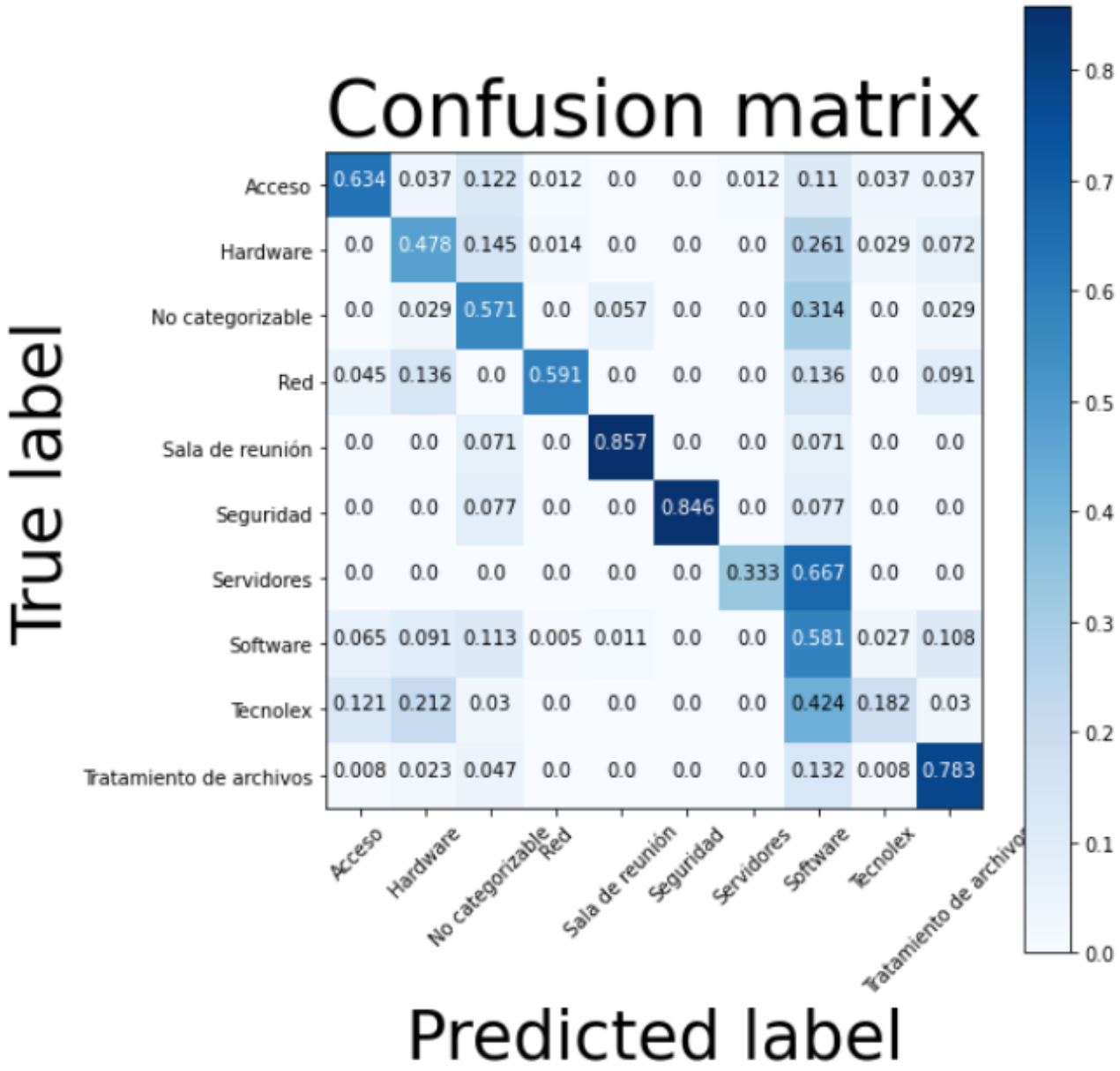

Figura 31: Matriz de confusión normalizada para el Árbol de clasificación

Anexo I: Classification Report para atributo Tipo Requerimiento de la cuarta base de datos

		Regresión	SVM	Árbol Decisión
Consulta	Precision	0.65	0.62	0.62
	Recall	0.62	0.58	0.58
	F1-Score	0.63	0.60	0.60
	Support	1609	1609	1609
Incidente	Precision	0.66	0.66	0.66
	Recall	0.71	0.72	0.72
	F1-Score	0.68	0.69	0.69
	Support	1610	1610	1610
Requerimiento	Precision	0.65	0.62	0.62
	Recall	0.64	0.59	0.59
	F1-Score	0.64	0.60	0.60
	Support	1610	1610	1610

Tabla 7: Classification Report para los modelos aplicados a Tipo Requerimiento

Anexo J: Matrices de confusión normalizada para atributo Tipo Requerimiento de la cuarta base de datos

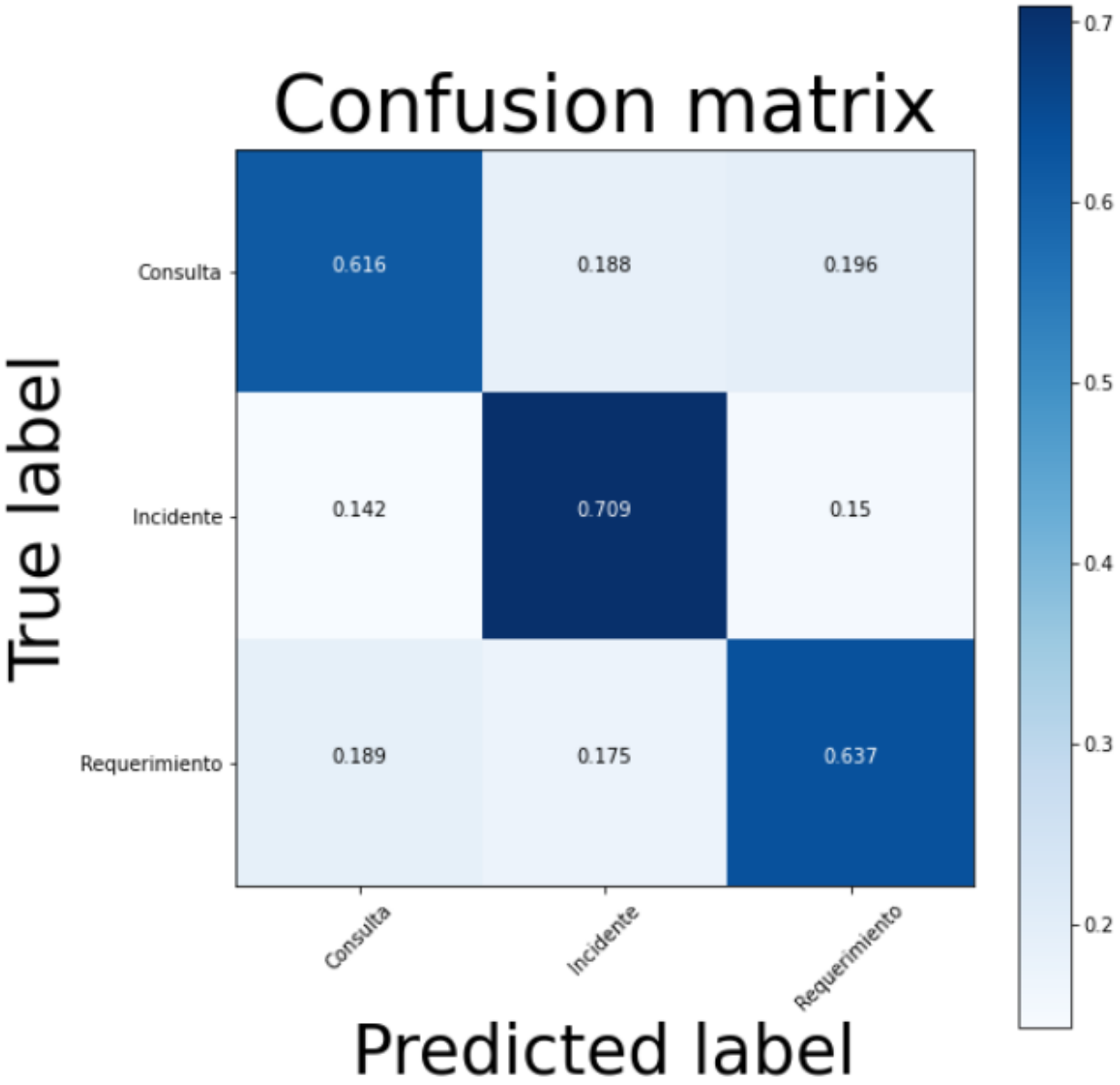

Figura 32: Matriz de confusión normalizada para la Regresión Logística

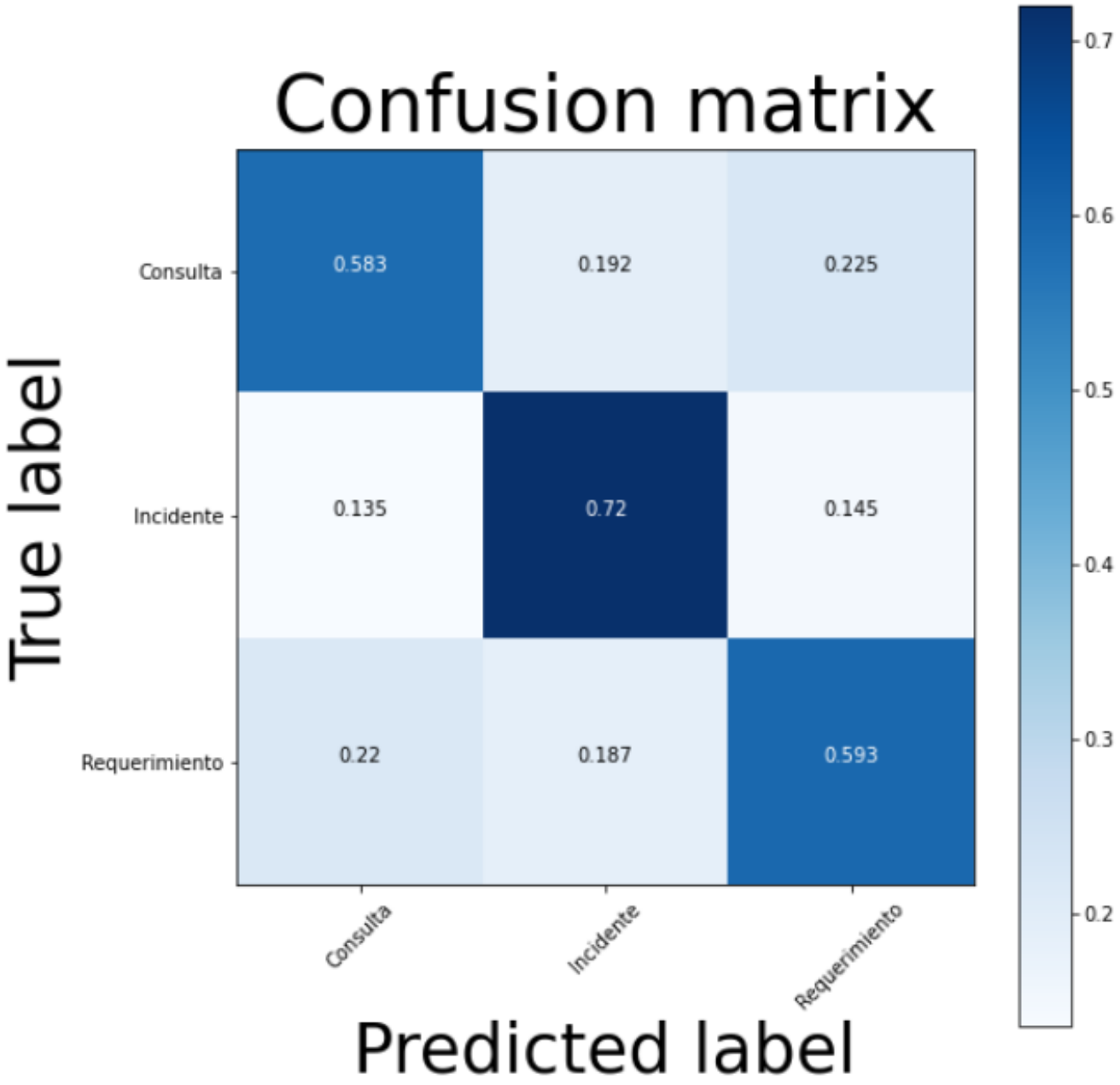

Figura 33: Matriz de confusión normalizada para el SVM

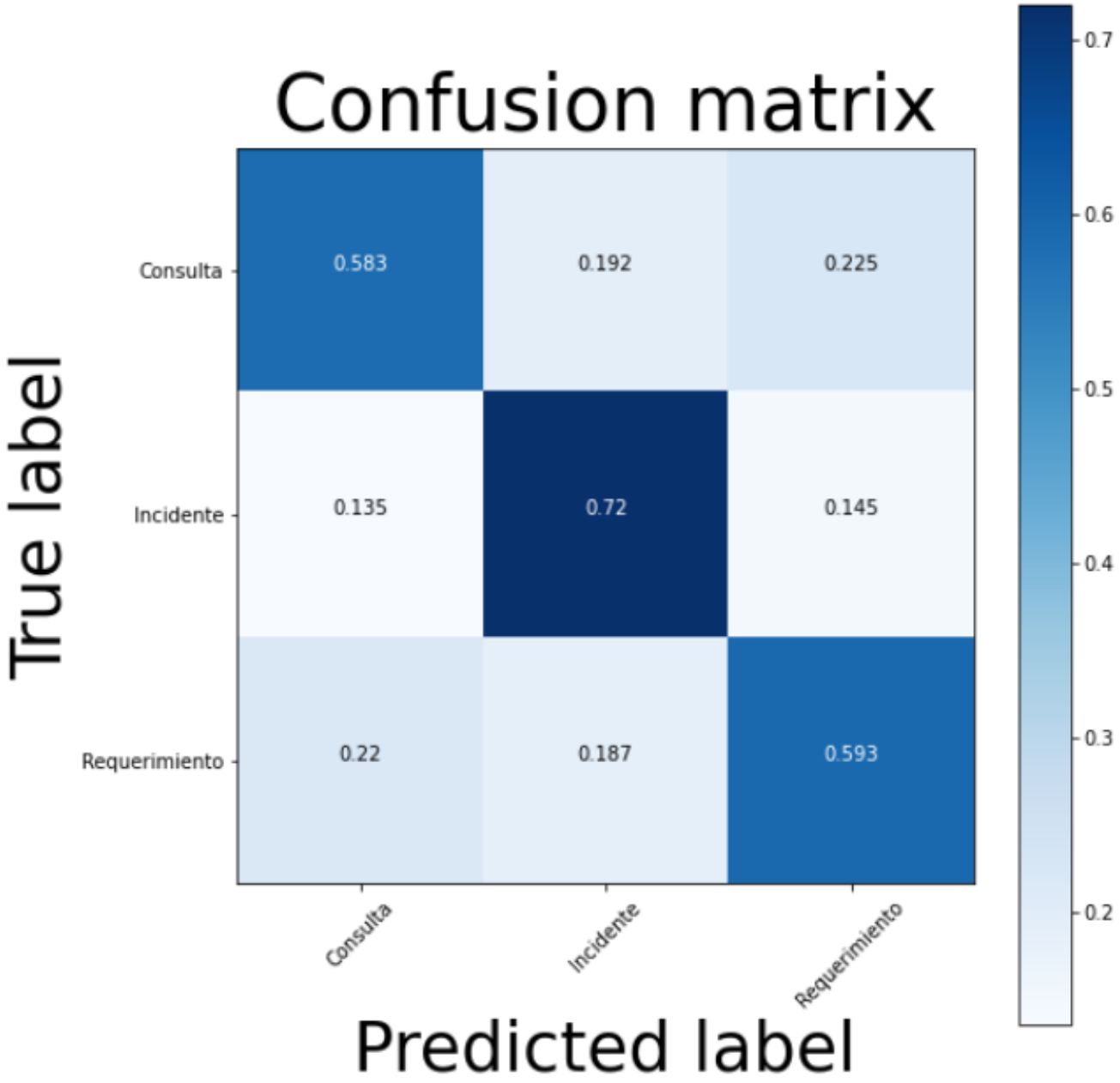

Figura 34: Matriz de confusión normalizada para el Árbol de clasificación

Anexo K: Classification Report para atributo Categoría de la quinta base de datos

		Regresión	SVM	Árbol Decisión
Administración y Comunicación	Precision	0.82	0.76	0.72
	Recall	0.70	0.72	0.69
	F1-Score	0.76	0.84	0.70
	Support	1748	1748	1748
Hardware y Redes	Precision	0.78	0.76	0.69
	Recall	0.71	0.70	0.66
	F1-Score	0.74	0.73	0.67
	Support	4496	4496	4496
Seguridad y Acceso	Precision	0.77	0.74	0.68
	Recall	0.68	0.67	0.63
	F1-Score	0.72	0.70	0.65
	Support	2620	2620	2620
Software	Precision	0.74	0.73	0.70
	Recall	0.83	0.82	0.75
	F1-Score	0.78	0.78	0.72
	Support	9099	9099	9099
Tratamiento de archivos	Precision	0.83	0.84	0.77
	Recall	0.80	0.79	0.76
	F1-Score	0.82	0.81	0.77
	Support	5404	5404	5404

Tabla 8: Classification Report para los modelos aplicados a Categoría

Anexo L: Matrices de confusión normalizada para atributo Categoría de la quinta base de datos

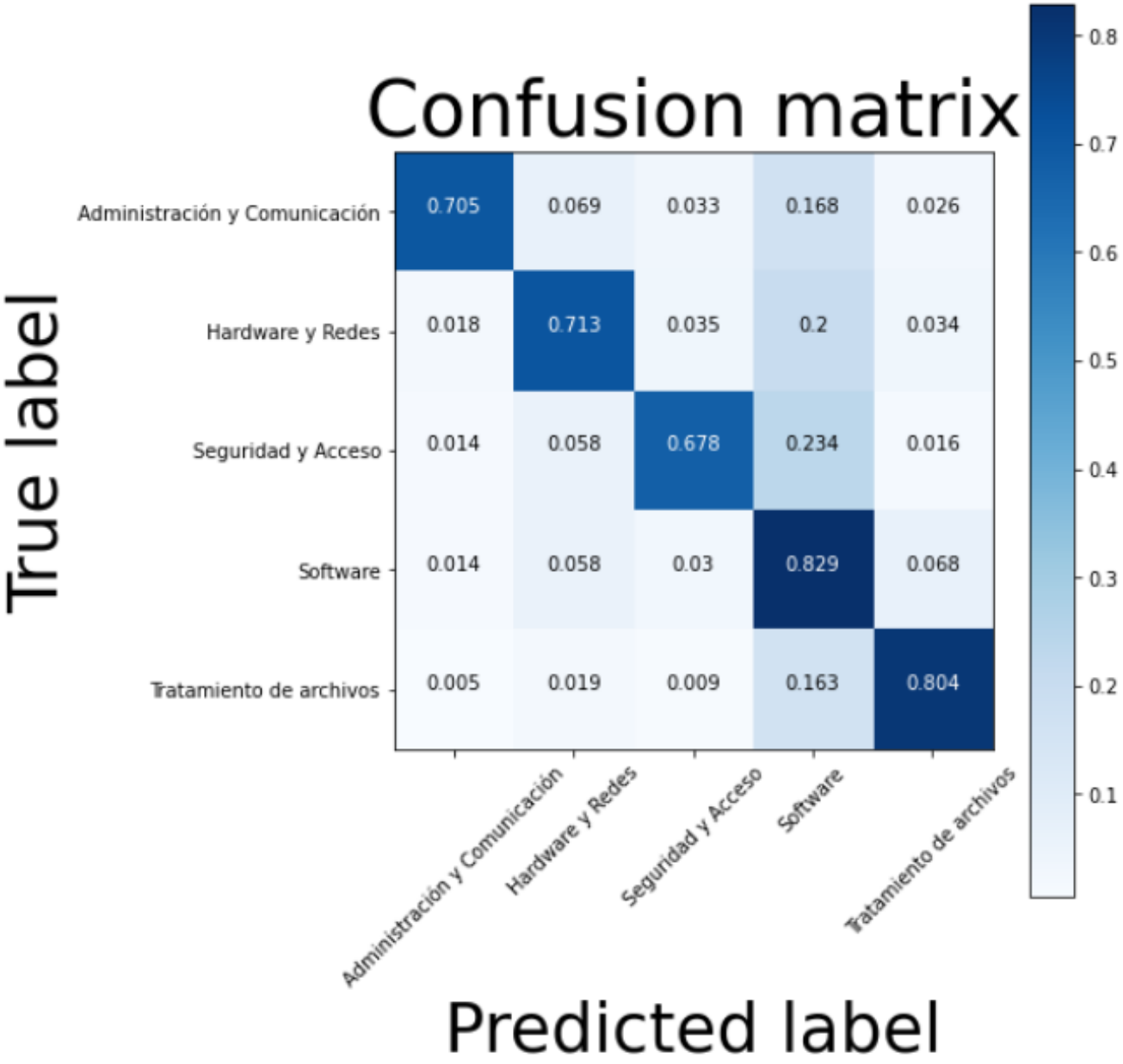

Figura 35: Matriz de confusión normalizada para la Regresión Logística

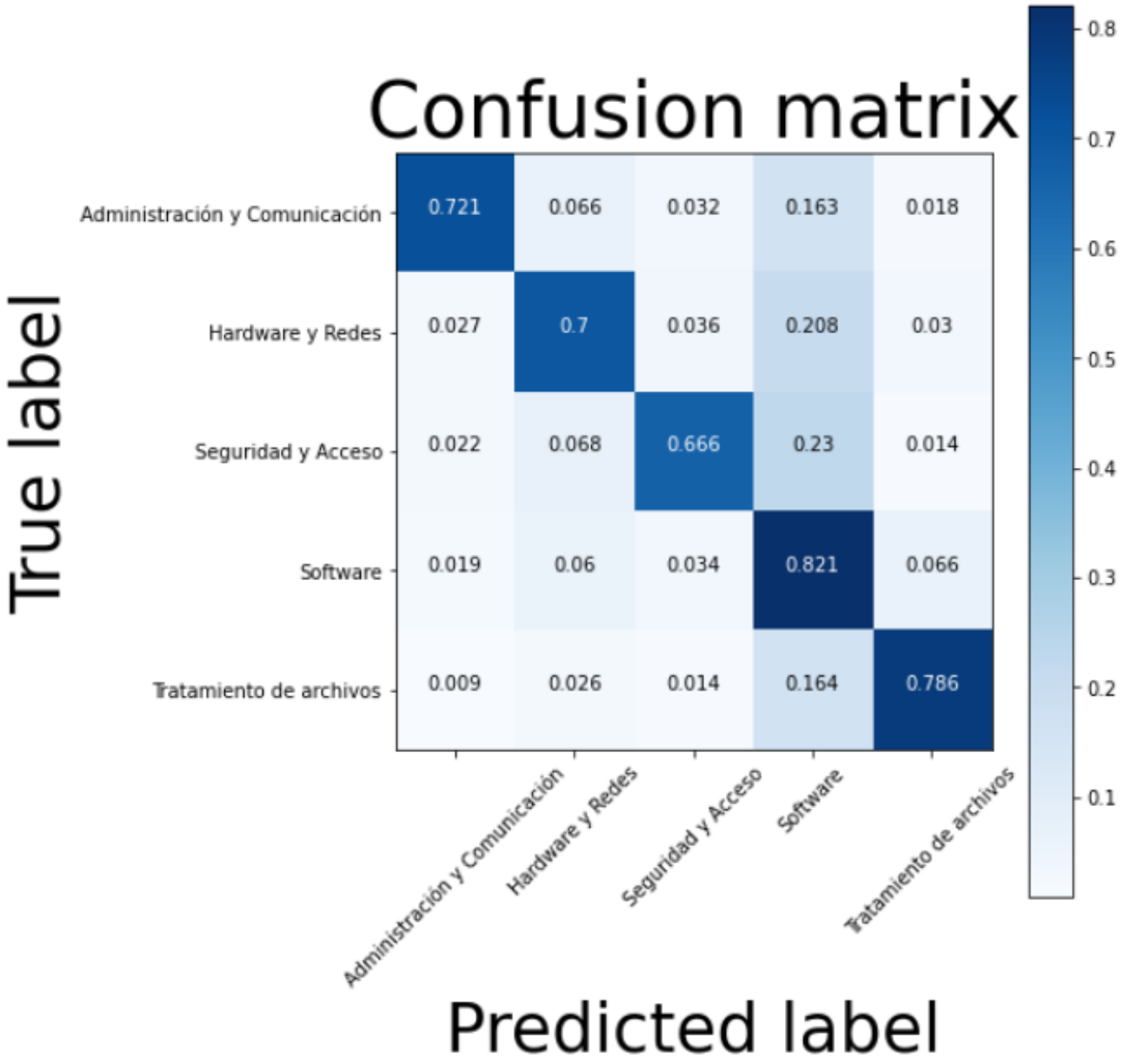

Figura 36: Matriz de confusión normalizada para el SVM

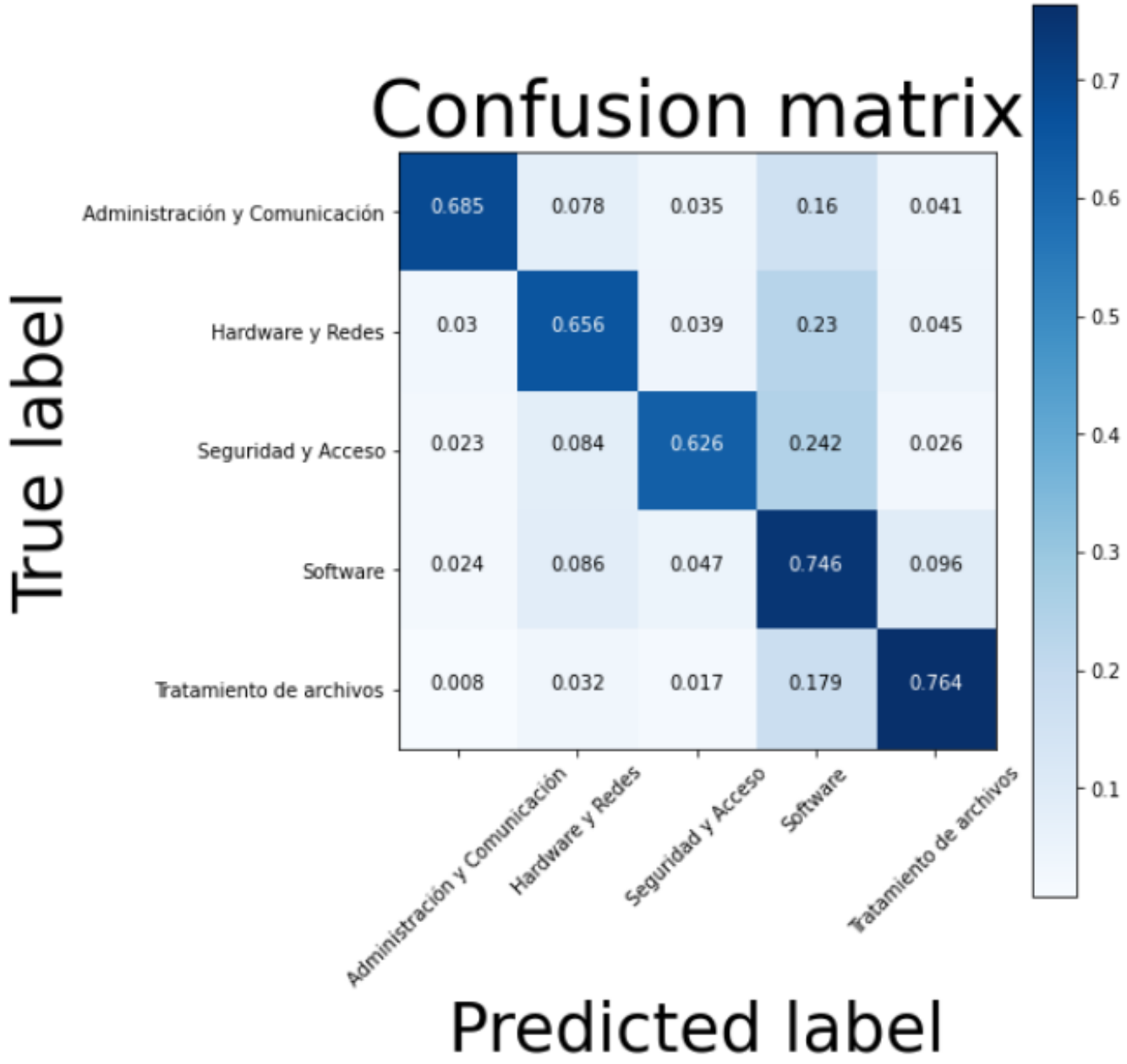

Figura 37: Matriz de confusión normalizada para el Árbol de clasificación