ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELECTRÍCA Y ELECTRÓNICA

EVALUACIÓN DE PRÁCTICAS DE PRIVACIDAD EN APLICACIONES MÓVILES

DESARROLLO DE UN MÓDULO DE ETIQUETADO DE PRÁCTICAS DE TRANSFERENCIA DE DATOS PERSONALES EN POLÍTICAS DE PRIVACIDAD EN ESPAÑOL USANDO TÉCNICAS PLN Y APRENDIZAJE AUTOMÁTICO

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

DARÍO JAVIER CASAGALLO AMAGUAYA

dario.casagallo@epn.edu.ec

DIRECTOR: DANNY SANTIAGO GUAMÁN LOACHAMIN

danny.guaman@epn.edu.ec

DMQ, MARZO 2023

CERTIFICACIONES

Yo, Darío Javier Casagallo Amaguaya declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

DARÍO JAVIER CASAGALLO AMAGUAYA

Certifico que el presente trabajo de integración curricular fue desarrollado por Darío Javier Casagallo Amaguaya, bajo mi supervisión.

DANNY SANTIAGO GUAMÁN LOACHAMIN
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

DARÍO JAVIER CASAGALLO AMAGUAYA

DANNY SANTIAGO GUAMÁN LOACHAMIN

DEDICATORIA

El presente trabajo va dedicado a Dios, que siempre me brindó fortaleza y salud para seguir adelante en todo este camino.

A mis padres Iván y Azucena, a mi hermana Belén y mi abuelita Blanca que siempre me apoyaron e inculcaron el valor de la responsabilidad; su crianza se refleja en todo lo que se ha luchado para alcanzar esta meta.

A mi compañera de vida, Gina que siempre estuvo presente en cada momento de mi vida, desde un inicio me apoyaste en todo y no me dejaste caer; tus sabios consejos y amor me ayudaron y dieron el valor moral suficiente para alcanzar esta meta.

A mis abuelitos que siempre me guiaron en el camino de vida, a ustedes Luis Casagallo (+), Luz Carlosama (+), Francisco Amaguaya (+) que me quisieron ver profesional, pero se adelantaron al llamado de Dios durante este camino, esto es para ustedes.

AGRADECIMIENTO

Agradezco primeramente a Dios por permitirme terminar mis estudios en esta prestigiosa universidad y por darme la sabiduría, salud y fortaleza necesaria durante este camino.

A mis padres y hermana por el apoyo brindado, sin duda alguna ustedes son pilares fundamentales en mi vida y esta etapa académica.

A mi compañera de vida, Gina, que siempre estuvo ahí para apoyarme, gracias a tus sabios consejos. Hoy lo estamos logrando.

Al PhD. Danny Santiago Guamán, que ha sido un pilar fundamental en mi desarrollo académico estos últimos semestres; gracias por su apoyo y colaboración, sobre todo en la realización del presente trabajo de titulación.

A cada docente de la Escuela Politécnica Nacional que supo compartir su conocimiento académico y espiritual durante mi transcurso educativo, a todos ellos muchas gracias por su ayuda y amistad incondicional.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE FIGURAS

RESUMEN

El incumplimiento de los requerimientos establecidos dentro de la Ley Orgánica de Protección de Datos Personales (LOPDP) puede llevar a fuertes sanciones a aquellas organizaciones involucradas. Existen requerimientos relacionados con transparencia, que persiguen informar adecuadamente a los usuarios sobre los tratamientos de datos personales. Las políticas de privacidad son el mecanismo de facto para proveer tal transparencia.

Con el objetivo de evaluar si una organización, o alguno de sus sistemas de software, cumple con lo estipulado en una política, se requiere extraer las prácticas de privacidad o tratamientos ahí definidos. No obstante, esta extracción no puede ser realizada manualmente, ya que no escalaría en los ecosistemas digitales actuales caracterizados por tener una inmensa cantidad de dispositivos (y políticas).

Para continuar con los esfuerzos por contribuir a la provisión de técnicas y herramientas para la evaluación automática de cumplimiento de requerimientos de privacidad y protección de datos, este trabajo presenta un módulo de etiquetado de prácticas de transferencias de datos en políticas de privacidad en español. Para ello, se hace uso de algoritmos de aprendizaje automático y de procesamiento de lenguaje natural.

En el capítulo 1 se presenta el fundamento teórico relacionado con la privacidad y protección de datos, y con las técnicas de aprendizaje automático y procesamiento de lenguaje natural, que son necesarios para comprender el presente documento. Además, se detallan algunas herramientas y lenguajes de programación utilizados durante el desarrollo de este trabajo.

En el capítulo 2 se presenta la metodología utilizada para el desarrollo del módulo de etiquetado de prácticas de transferencias de datos en políticas de privacidad en español. Además, se presentan los códigos utilizados para construir los clasificadores basados en aprendizaje automático, para realizar pruebas y para determinar los modelos con el mejor rendimiento. Se construye un ensamble con los modelos seleccionados y se calcula ciertos valores para evaluar el rendimiento de los modelos elaborados.

Finalmente, en el capítulo 3 se realiza un análisis de los resultados obtenidos durante el proceso, y se presentan las conclusiones y recomendaciones, una vez que se ha finalizado el trabajo de integración curricular.

PALABRAS CLAVE: privacidad, transferencia de datos, protección de datos, algoritmos supervisados, clasificación, machine learning, lenguaje natural, políticas de privacidad

ABSTRACT

Failing to comply with requirements set out in the Ley Orgánica de Protección de Datos Personales (LOPDP) may lead penalties to the involved organizations. There exist transparency-related requirements, which seek to inform users on personal data processing. Privacy policies are the main mechanism to provide such a transparency.

To assess compliance with statements laid down in privacy policies, privacy practices firstly need to be extracted. Yet, such extraction cannot be manually conducted, since it would not scale in the current digital ecosystem, which have a vast number of devices (and policies).

To continue with the efforts to contribute to the provision of techniques and tools for compliance assessment with privacy and data protection requirements, this work presents a module for tagging data transfer practices in Spanish privacy policies, using machine learning and natural language processing algorithms.

Chapter 1 presents the theoretical background related to privacy and data protection, and to machine learning and natural language processing techniques, which are necessary to understand this document. Furthermore, it details the tools and programming languages used in this work.

Chapter 2 presents the methodology used to develop the module for tagging data transfer practices in Spanish privacy policies. In addition, the codes used to build the machine learning-based classifiers, to perform tests and to determine the models with the best performance are presented. An ensemble is built with the selected models and standard metrics are calculated to evaluate the performance of the elaborated models.

Finally, Chapter 3 analyses the results obtained during the process and presents the conclusions and recommendations once the project has been completed.

KEYWORDS: privacy, data transfer, data protection, algorithms, supervised, classification, machine learning, natural languages, privacy policies

1 INTRODUCCIÓN

Dentro de la Ley Orgánica de Protección de Datos Personales (LOPDP) se establece un conjunto de principios y requisitos que las organizaciones, y los sistemas de software, tienen que cumplir cuando tratan datos personales. El incumplimiento de dichos reglamentos puede conllevar fuertes sanciones de un cierto porcentaje de su facturación [1].

La ley en si surge como un mecanismo jurídico para proteger el derecho a la privacidad de las personas en el área tecnológica. Sus objetivos principales son: i) definir los datos personales ii) determinar el responsable del tratamiento de datos iii) regular cuestiones esenciales del tratamiento de datos, tales como la conservación, el acceso, la seguridad y la confidencialidad y iv) determinar el nivel de protección adecuado para la transferencia de datos personales a otros países [2].

La LOPDP define las políticas de privacidad como un medio para proveer transparencia en el tratamiento de datos, pues establece que los usuarios tienen que ser informados sobre las prácticas de privacidad que se van a llevar a cabo (p.ej., recolección, transferencia, almacenamiento, etc.). Sin embargo, la evaluación de la transparencia requiere identificar múltiples elementos de cada práctica de privacidad [1]. Por ejemplo, para una práctica de transferencia de datos se requiere identificar el tipo de dato a transferir, el país al cual se comparte, el mecanismo de transferencia, etc.

Hoy en día existen enfoques basados en algoritmos de aprendizaje automático (conocido también como Machine Learning) y técnicas de procesamiento de lenguaje natural (PLN) que permiten identificar ciertos elementos de prácticas de transferencia de datos en políticas de privacidad textuales en inglés. Sin embargo, no existen muchas las contribuciones para identificar ciertos elementos de prácticas de transferencia de datos en políticas de privacidad textuales en español son escasas, lo que dificulta la evaluación del cumplimiento de requisitos de transparencia establecidos dentro de la LOPDPD ecuatoriana [1].

En este contexto, para contribuir con la evaluación automatizada del cumplimiento de transparencia en el tratado de datos, este proyecto se centra en desarrollar un modelo de clasificación basado en aprendizaje automático para etiquetar prácticas de transferencia de datos personales en políticas de privacidad textuales en español.

1.1 Objetivo general

Desarrollar un modelo de clasificación de prácticas de transferencia de datos personales en políticas de privacidad en español usando técnicas PLN y aprendizaje automático.

1.2 Objetivos específicos

- Estudiar los fundamentos teóricos relativos a las técnicas PLN y aprendizaje automático para el desarrollo del modelo de clasificación.
- Estudiar los fundamentos teóricos relativos a la privacidad y protección de datos, con énfasis en los requisitos de transparencia.
- Diseñar un protocolo para la elaboración de un dataset que será usado para construir el modelo de clasificación.
- Implementar el modelo de clasificación necesario para extraer las prácticas de transferencia de datos y sus parámetros.
- Validar el modelo de clasificación implementado.

1.3 Alcance

Dentro de las prácticas de privacidad, que se estipulan en la LOPDP, está la recolección, el almacenamiento, la transferencia, entre otros. El presente proyecto se enfocó en las prácticas de transferencia de datos. Concretamente, se construyó un modelo de clasificación basado en aprendizaje automático para etiquetar las prácticas de transferencia de datos en políticas de privacidad textuales en español.

Para su desarrollo se tomó en cuenta dos fases fundamentales: 1) generación del dataset y 2) desarrollo del modelo de clasificación adecuado para clasificar las prácticas de transferencia de datos dentro de políticas de privacidad en español. A continuación, se da un breve resumen del proceso que se llevó a cabo en cada una de las fases.

1. **Generación del dataset**

La Figura 1 muestra en resumen el proceso llevado a cabo para la generación del dataset, el cual consta de cinco pasos que permitieron obtener la información necesaria que fue usada para la creación del modelo de clasificación.

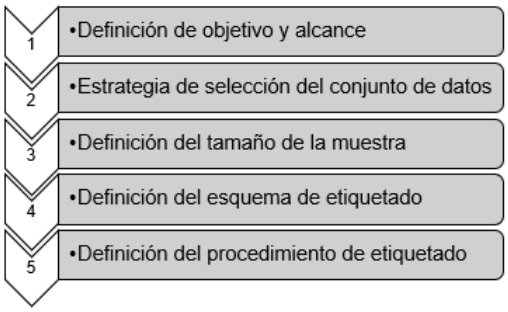


Figura 1.1. Proceso para generación de Dataset.

Para la generación del dataset se hizo uso de la herramienta “AnnoTool-CLIIP”; desarrollada en un proyecto anterior por Félix Miño [3], la cual nos ofreció un ambiente favorable para el desarrollo y creación del dataset necesario. En este caso se recolectó información específica de prácticas de transferencia de datos personales a través de ciertos elementos de transparencia como tipo de datos, especificación de una tercera entidad, mecanismo de transferencia, etc.

Desarrollo del modelo de clasificación

La Figura 2 por su parte muestra el proceso utilizado luego de la obtención del dataset, para desarrollar el modelo de clasificación. Este consta de siete pasos que permitieron la exportación de un modelo de clasificación adecuado para etiquetar las prácticas de transferencia de datos en políticas de privacidad en español.

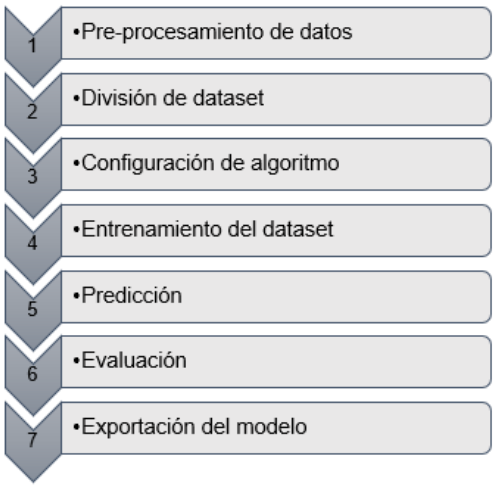


Figura 1.2. Proceso para elaboración de modelo.

Para el desarrollo del modelo de clasificación se hizo uso del lenguaje de programación Python que, junto con sus librerías Scikit learn y NLTK, nos proporcionó las funciones necesarias para el análisis de datos y la evaluación del modelo desarrollado. Se hizo uso de al menos dos algoritmos de aprendizaje automático para la construcción del modelo de clasificación (p.ej. regresión logística y SVM)

El modelo desarrollado fue evaluado a través de pruebas de funcionamiento, donde se comprobó el rendimiento del mismo en varios conjuntos de datos que contienen políticas de privacidad en español. Esto se evidenció con el uso de las distintas métricas estándar específicas para el modelo (p.ej. matriz de confusión).

Finalmente se obtuvo un modelo de clasificación de prácticas de transferencia de datos adecuado.

1.4 Marco teórico

En esta sección se presentan algunos conceptos relacionados a la privacidad y protección de datos, que son necesarios para comprender el contexto del presente trabajo. Se realiza un enfoque al concepto de privacidad como integridad contextual, según Nissenbaum. Además, se abordan tres perspectivas para el estudio de la privacidad y protección de datos en el ámbito de la ingeniería propuesto por Troncoso. Luego se detallan algunos puntos importantes de la Ley Orgánica de Protección de Datos Personales (LOPDP) y se da una definición de políticas de privacidad según las definiciones y principios de dicha ley. Así también se toma ciertas ideas de trabajos ya realizados para extraer una lista con prácticas de privacidad, las cuales son objeto de estudio. Finalmente, se detallan algunas herramientas y lenguajes de programación utilizados para la implantación de un modelo que permita verificar entre las políticas de privacidad aquellas que son de transferencia de datos.

1.4.1. Privacidad y protección de datos personales

Por muchos años la privacidad ha sido estudiada desde diferentes perspectivas, incluyendo el ámbito social, legal y de ingeniería, con dos objetivos principales: (i) entender de mejor manera que es privacidad en el contexto social y (ii) asegurar que los marcos legales den soporte a la privacidad como un derecho.

Para abordar de mejor manera este proyecto es necesario mencionar algunas definiciones, como privacidad y datos personales los cuales se detallan a continuación.

1.4.1.1. Privacidad

Actualmente, en el mundo se maneja gran cantidad de datos personales de los ciudadanos, que se ha ampliado con los avances tecnológicos. A la par la privacidad de los ciudadanos se ha visto amenazada. En la literatura existen varias definiciones de privacidad. No obstante, para los propósitos de este trabajo se define a la privacidad como integridad contextual. Este concepto fue provisto por Nissenbaum en 1997, quien plantea un marco conceptual en donde relaciona la protección de los datos privados a las normas específicas que imperan en cada contexto.

Este planteamiento explica que en cada espacio o entorno donde se desenvuelve el hombre existen normas explicitas o implícitas que controlan o limitan el comportamiento natural del ser humano. La idea central de la integridad contextual se fundamenta en la siguiente creencia:

> "*No hay arenas de la vida no gobernadas por normas informacionales [...] Prácticamente todo: las actividades que llevamos a cabo, los acontecimientos que*

suceden, las transacciones que realizamos... todo ocurre en un contexto no sólo en cuanto a lugar, sino que conllevan unas convenciones y expectaciones culturales" [4]

De acuerdo con esta definición, no es necesario idealizarnos normas que protejan la vida privada de cada uno, sino aceptar la existencia de ciertos dominios que están ya relacionados con la experiencia común y al mismo tiempo gobernados por normas específicas.

De igual manera se enfoca en que de cierta manera los datos personales son compartidos, pero estrictamente regidos por las normas que controlan el flujo de información personal, que indican hasta qué punto un individuo puede hacer visible su información privada y cuando nuestra vida privada es invadida. Dado que el marco de protección de datos privados es la integridad contextual, un incumplimiento de las normas que regulan el contexto supondría una violación a la privacidad, la cual podría ser justificada por una fuerza mayor, tal como lo menciona Nissenbaum: El derecho a la intimidad y vida privada "*no es un derecho al secreto, ni al control de la información, sino un derecho al apropiado flujo de información personal*" [5]

En este sentido se menciona que la información debe alinearse a los flujos de información personal permitidos, de manera que una organización no llegue a afectar la privacidad de una persona.

1.4.1.2. Datos personales

La Ley Orgánica de Protección de Datos Personales (LOPDP) define a los datos personales como cualquier dato que "*identifica o hace identificable a una persona natural, directa o indirectamente, en el presente o en el futuro. Los datos inocuos, metadatos o fragmentos de datos que identifiquen o hagan identificable a un ser humano, forman parte de este concepto*" [1].

En la definición presentada por la LOPDP existen ciertos puntos importantes que permiten tener una idea más clara de lo que son los datos personales, los cuales se presentan a continuación:

- Sujeto de protección: Aplica solo para personas naturales (p. ej., usuarios).
- Dato que identifica: Tipo de dato que identifica directamente a una persona (p.ej., CI, ADN, Huella dactilar).
- Dato que hace identificable: Tipo de dato que no identifica directamente a una persona, sin embargo, en conjunto con otros datos puede llegar a hacerlo (p. ej., dirección, nombre, número telefónico).
- Metadatos: Se refiere al uso de identificadores para rastrear a una persona o construir un perfil de esta (p. ej., estatura, nombre, color de cabello, color de piel, etc.).

Hoy en día el uso de datos personales en aplicaciones móviles es muy común, por lo que, preservar su confidencialidad dependerá de la tecnología y seguridad que maneje cada organización dentro de la aplicación.

1.4.2. Tecnologías para protección de datos personales

La recopilación de datos personales en aplicaciones y sitios web se ha hecho muy frecuente en los últimos años. La protección de la información es una pieza clave para el funcionamiento de sistemas que manejan gran cantidad de datos personales. Algunos de estos manejan ciertos protocolos y estándares de seguridad para precautelar la privacidad de las personas. Sin embargo, en ocasiones no es suficiente, dado que, por ejemplo, los datos enviados por la red pueden ser interceptados y publicados en distintas plataformas con el fin de sacar algún provecho sin importar los problemas físicos y psicológicos causados a los usuarios víctimas de la denominada violación de su privacidad.

La compartición de datos personales a veces es necesaria, ya que en ocasiones es indispensable la recolección de los mismos para la provisión de un servicio, claro que sujetas a las normas de cada contexto, tal como lo menciona Nissenbaum. Si lo flujos de información no se ajusta a las normas definidas en un contexto entonces se produce una violación a la privacidad [4]. Para mitigar el riesgo de una violación a la privacidad, Carmela Troncoso define 3 tecnologías de protección que se detallan a continuación.

1.4.2.1. Privacidad como confidencialidad

Las tecnologías bajo este paradigma se alinean con el derecho a estar solo, que busca evitar que la información personal de los usuarios sea accesible para cualquier persona o entidad. En otras palabras, se preocupa por ocultar la información a un público más amplio que el deseado. El objetivo de este tipo de tecnología es permitir el uso de servicios mientras se minimiza la cantidad de información expuesta. Para esto se presentan dos enfoques que se relacionan a la protección de datos, el primero es haciendo uso de primitivas criptográficas que nos permitan la inferencia de datos (p.ej., claves simétricas, claves públicas, hash) y el segundo el uso de métodos que controlen la divulgación de estos (p.ej., políticas de privacidad).

1.4.2.2 Privacidad como control

Las tecnologías bajo este paradigma buscan el control de la divulgación de los datos. Aquí se asume que ciertos datos necesariamente tienen que ser divulgados a los proveedores para la provisión de un servicio. Ante esto, la privacidad como control requiere abordar dos aspectos: (1) permitir que el usuario determine cómo pueden hacer uso de sus datos y (2)

permitir que las organizaciones apliquen las políticas informadas a los usuarios para así evitar el uso indebido de la información. Para esto se definen tres técnicas:

- Soporte para configuración de privacidad: Son aquellos controles de un servicio que permite a los usuarios expresar como quieren que se manejen sus datos.
- Soporte para negociación de políticas de privacidad: Uso de tecnologías para la configuración de políticas de privacidad entre el usuario y el proveedor de servicios.
- Soporte para la interpretabilidad de la política de privacidad: Actualmente se tiene dos enfoques para comprender las políticas de privacidad: (1) el uso de etiquetas e interpretación y análisis por parte de expertos y (2) la automatización del proceso de interpretación.

1.4.2.3. Privacidad como transparencia

Las tecnologías bajo este paradigma buscan la provisión de información sobre el tratamiento de los datos personales de un usuario. De este modo, el usuario puede ser consciente de los potenciales riesgos a su privacidad. Aquí se abordan dos mecanismos que analizan la actividad de usuario frente al flujo y tratamiento de datos:

- Transparencia basada en avisos: El objetivo es usar avisos, que orienten al usuario sobre los riesgos a su privacidad. Un mecanismo en esta categoría son los espejos, en el cual se da a conocer como los demás ven nuestros datos.
- Transparencia basada en auditoria: Realizar este tipo de auditoria requiere de un registro de todas las operaciones de acceso y procesamiento de datos del sistema, con lo cual se puede verificar, por ejemplo, cuando se transmiten los datos a otros.

1.4.3. Evaluación de cumplimiento de privacidad y protección de datos

Dentro de la Ley Orgánica de Protección de Datos Personales (LOPDP) se ha definido un conjunto de requisitos que deben ser cumplidos por las organizaciones que procesan datos personales. En caso de incumplimientos, se establecen ciertas penalizaciones. En la siguiente sección se presentan los aspectos más relevantes de la LOPDP, haciendo énfasis, para los propósitos de este trabajo, en los requisitos de transparencia y en los mecanismos de evaluación más relevantes.

1.4.3.1. Ley Orgánica de Protección de Datos Personales (LOPDP)

La Asamblea Nacional del Ecuador, en su sesión virtual No. 707 del 10 de mayo de 2021, aprobó el proyecto de Ley, posteriormente el 26 de mayo de 2021 se publica la Ley Orgánica de Protección de Datos Personales (LOPDP) que tiene como objetivo garantizar el derecho

a protección de datos personales, que incluye el acceso y decisión sobre la información y datos de este carácter, así como su correspondiente protección.

Principalmente la ley hace hincapié en las condiciones que se deben verificar para que el tratamiento de los datos personales sea legítimo, además, de las formas por las cuales a través el titular de la información pueda manifestar su voluntad para el tratamiento de sus datos. Además, regula el contenido y alcance de los derechos:

- A la información
- De acceso
- De rectificación y actualización
- De eliminación de oposición
- A la portabilidad
- A no ser objeto de una decisión basada en valoraciones automatizadas
- De consulta pública y gratuita ante el Registro Nacional de Protección de datos personales
- A la educación digital

Por otro lado, brinda ciertos conceptos de categorías de datos personales, como los datos sensibles, los de niños, niñas y adolescentes, los de salud y los de las personas con discapacidad; y se centra en el tratamiento especializado de estos datos.

Por último, la ley establece ciertas sanciones para aquellos que incumplan las normas relacionadas al tratamiento de datos personales. Así destacamos que la LOPDP busca la protección de ciudadano y no de las empresas al reconocer en sus principios básicos, en casos de duda, la aplicación favorable al titular.

1.4.3.2. Políticas de privacidad

Las políticas de privacidad, también conocidas como políticas de protección de datos en la LOPDP, son un mecanismo para informar al usuario la manera en que un servicio hace uso de sus datos personales. De acuerdo con la LOPDP "*el responsable del tratamiento de datos debe tomar las medidas apropiadas para proporcionar a sus usuarios cualquier información en relación con el procesamiento de datos, en una forma concisa, transparente, inteligible y de fácil acceso, utilizando un formato claro e idioma sencillo, la información se proporcionará por escrito o por otros medios, esta información deberá ser comunicada el momento mismo de la recogida del dato personal*" [1]. En este contexto, las políticas de privacidad se han convertido en el mecanismo de facto de transparencia que una organización usa para informar a los usuarios sobre los tratamientos o prácticas de privacidad que llevará a cabo.

.

1.4.3.3. Prácticas de privacidad

Dentro de los estatutos de la LOPDP se definen ciertos requisitos que deben cumplir aquellos responsables del tratamiento de datos. Los tratamientos de datos por lo general tienen una estructura específica que es:

1. Tratamiento
1.1. Atributo
1.1.1. Valor

Félix Miño, en su trabajo de titulación menciona que "*Estos tratamientos no están explícitamente expresados en la PLOPDP, por lo que tienen que extraerse de los distintos artículos que componen esta ley*" [3]

Para estos tratamientos se menciona que existen ciertas categorías, las cuales se expresan por un sustantivo derivado de un verbo, por ejemplo, recolección. Para los propósitos de este trabajo, a estos tratamientos los denominaremos prácticas de privacidad. Los atributos por su parte se los considera como componentes de un tratamiento de datos y ayudan a que estos sean entendidos de mejor manera. Por ejemplo, el "tipo de dato personal" o la "finalidad" son potenciales atributos de una práctica de recolección. Por último, el valor que tiene cada atributo es una descripción aún más a detalle de este. Por ejemplo, la localización, los identificadores o información demográfica son potenciales valores de un atributo "tipo de dato personal".

Las principales prácticas de privacidad y atributos identificados en la LOPDP son:

Recolección: se menciona en el artículo 23 de LOPDP, así como sus atributos: tipo de dato recolectado, la finalidad o propósito de la recolección, el origen de los datos y la elección del usuario en cuanto a esta recolección [1].

Conservación: se lo menciona en el artículo 17 de la LODPD, así como sus atributos: periodo de conservación y propósito de conservación [1].

Transferencia: se menciona en el artículo 23, este tratamiento es el que más se explica a detalle en la LOPDP. Sus atributos son: transferencia a terceros, transferencia internacional y propósito de la transferencia. Félix Miño menciona que:

> "*La principal diferencia entre la transferencia a terceros y la transferencia internacional es que la transferencia internacional de datos se refiere a que los datos son enviados fuera del territorio nacional, pero el responsable del tratamiento de datos sigue siendo el mismo que en un inicio, mientras que la transferencia a terceros la ubicación geográfica del destinatario puede estar o no en el territorio nacional…*" [3] [1]

Es por esto que en cada uno de sus artículos la LOPDP explica de mejor manera la transferencia a terceros y la internacional, dando a cada uno de estos los artículos suficientes para su mejor entendimiento.

Control del titular: de la misma manera se lo menciona en el artículo 23 de la LOPDP. Sus atributos son: acceso, edición y eliminación. Estos atributos están definidos en los artículos 24, 25 y 26, respectivamente [1].

Las prácticas de privacidad con sus respectivos atributos y valores definidos en la LOPDP se presentan en la Tabla 1.1.

Tabla 1.1. Resumen de prácticas de privacidad.

Práctica	Atributo	Valor
Recolección	Tipo de dato	Genérico
	Propósito	Genérico
	Origen de los datos	Genérico
	Elección del usuario	Genérico
Conservación	Periodo de conservación	Genérico
	Propósito	Genérico
Transferencia	Transferencia internacional	Genérico
	Transferencia a terceros	Genérico
	Propósito	Genérico
Control del titular	Acceso	Genérico
	Edición	Genérico
	Eliminación	Genérico

Los valores identificados en la Tabla 1.1 son los que se usan en la herramienta de anotación para la identificación de las políticas de privacidad del presente proyecto.

1.4.3.4. Extracción de prácticas de privacidad

En muchos casos, la divulgación de información puede ser inevitable, por lo que el uso de tecnologías que permitan minimizar el riesgo a la violación de privacidad es necesario para controlar el uso de la información. La divulgación de las prácticas de privacidad a través de políticas de privacidad permite que tanto el usuario como la organización conozcan cómo se tratarán los datos personales.

En este contexto, la interpretabilidad de las políticas de privacidad es importante. Para ello, se han desarrollado tecnologías que mejoran la capacidad de los usuarios para interpretar dichas políticas. Existen dos enfoques para mejorar la comprensión de las políticas de privacidad: la primera consiste en confiar en expertos que etiqueten, analicen y proporcionen razones para las políticas de privacidad existentes y la segunda consiste en automatizar el proceso de interpretación. Este trabajo se circunscribe en el segundo enfoque.

1.4.4. Procesamiento de lenguaje natural

"El procesamiento de lenguaje natural (PLN) involucra una transformación a una representación formal, manipula esta representación y, por último, si es necesario lleva todo nuevamente al lenguaje natural" [6]. Es así que el PLN se usa para la recuperación y extracción de la información, traducción automática, sistemas de búsquedas de respuestas, generación de resúmenes automáticos, minería de datos, análisis de sentimientos, entre otras.

A continuación, se presenta las diferentes operaciones usadas para el procesamiento de lenguaje natural.

1.4.4.1. Preprocesamiento

El preprocesamiento de datos es una técnica de minería de datos comúnmente utilizada para transformar datos recolectados sin procesar, en datos usables o en un formato adecuado para su respectivo análisis [7]. Dentro de las técnicas para el preprocesamiento de datos están:

- Limpieza de datos: Necesaria para corregir partes irrelevantes y faltantes en los datos. Implica el manejo de datos faltantes, datos ruidosos, etc.
- Transformación de datos: Transforma los datos en formas apropiadas para el proceso de minería. Esto implica la normalización de los datos, la selección de atributos, la discretización y la generación de jerarquías.
- Reducción de datos: su objetivo es aumentar la eficiencia del almacenamiento y reducir los costos de almacenamiento y análisis de datos. Para la reducción de datos se realizan los siguientes pasos: Agregación del cubo de datos, selección de subconjunto de atributos, reducción de numerosidad y reducción de dimensionalidad.

1.4.4.2. Vectorización

La vectorización de datos es una metodología muy usada para el análisis y la extracción de datos de manera estructurada desde un mapa digitalizado. Tal como lo dice su nombre, se basa en almacenar la información en un formato vectorial a través de ciertas primitivas que

la adecuan para su utilización en aplicaciones del tipo proporcionado por el sistema de información geográfico (GIS) [8].

Las primitivas más usadas para la vectorización son la definición de n-gramas y el recuento de frecuencias, las cuales se detallan a continuación.

- **Definición de n-gramas**

Para un mayor entendimiento de la vectorización mediante n-gramas se presentan ejemplos para la frase "*Transferencia de datos personales fuera del EEE*"

  - Ejemplo de características de 1-grama: *"transferencia", "personal", "datos", "fuera", "EEE"*.
  - Ejemplo de características para 2-gramas: *"transferencia personal", "datos personales", "datos fuera", "fuera del EEE".*
  - Ejemplo de características de 1-2 gramas: *"transferencia", "personal", "datos", "fuera", "EEE", "transferencia personal", "datos personales", "datos fuera", "fuera del EEE".*

En este sentido el n-grama que se deba utilizar depende del propio estudio y también se lo puede elaborar manualmente [8].

- **Recuento de frecuencias de las características**

El recuento nos dice como pueden ser identificados los segmentos:

  - Un documento es representado como un vector con valores numéricos: $(W_1, W_2, W_3, \dots, W_n)$
  - Binario:

$W_i = 1$ si el término i correspondiente se encuentra en el documento

$W_i = 0$ si el término i no está en el documento

  - Frecuencia de términos (TF):

$W_i = tfi$ donde tfi es el número de veces que el termino aparece en el documento.

  - Frecuencia inversa del documento (TF-IDF):

$W_i = tfi * idfi * \log\left(\frac{N}{dfi}\right)$ donde dfi es el número de documentos que contienen el termino i, y N el número total de documentos de la colección [8].

Así mismo, la aplicación de estos n-gramas depende del estudio que se esté llevando a cabo.

Todas estas premisas nos permiten llevar a cabo la vectorización con el fin de tener nuestra información estructurada y lista para ser analizada.

1.4.5. Aprendizaje automático

Por años el estudio del comportamiento humano ha sido de gran interés por quienes quieren verificar como el hombre ha ganado experiencia con el transcurso de los años. En este contexto se han desarrollado distintos estudios y experimentos con el objetivo de simular el aprendizaje del hombre en su entorno, lo que ha llevado a la implementación de ciertas metodologías que usan ciencias exactas como la matemática y estadística. Estas ayudan a generar un modelo de aprendizaje con los datos recopilados de varios estudios y tener una solución a un cierto problema; es aquí donde nace la expresión de aprendizaje automático, la cual hace referencia a la adquisición de conocimiento y a la innovación de este según la experiencia suscitada dentro de un entorno informático.

El aprendizaje automático puede ser un tema muy general y como se conoce es muy utilizado en lo que es la inteligencia artificial (AI). De aquí se despliegan ciertos conceptos que ayudan a entenderlo de mejor manera. El tipo de aprendizaje supervisado y el no supervisado se encuentran dentro de la rama de la IA, cada uno con sus respectivos modelos y metodologías que ayudan a la adquisición de un nuevo conocimiento, por lo cual también ayudan a simular el aprendizaje del ser humano en su interacción con los eventos y objetos de su entorno, gracias a esto hoy en día se tiene el reconocimiento de imágenes, el posicionamiento de una web en un motor de búsqueda y algo tan sonado como lo son los coches autónomos [9].

1.4.5.1. Aprendizaje no supervisado

También conocido como aprendizaje sin clase en el cual se manejan datos no etiquetados o sin una estructura fija, es decir, aquellos datos de los cuales no se tiene mucha información y por ende no se puede identificar un patrón o saber de qué clase son parte. Básicamente este aprendizaje se encarga de clasificar los datos creando subconjuntos partiendo de la exploración de un conjunto de datos indicado del cual se toma en cuenta las características similares de cada uno para agruparlos [10].

1.4.5.2. Aprendizaje supervisado

Conocido como aprendizaje con clase cuyo objetivo es obtener un patrón a partir de un conjunto de datos de entrenamiento, y que nos permita realizar predicciones en otros conjuntos de datos desconocidos. De aquí viene el término “supervisado” dado que se refiere al uso de datos de ejemplo como entrada para entrenar a un modelo del cual sabemos la salida o clase deseada.

Existen dos tipos principales de aprendizaje supervisado; regresión y clasificación. Los cuales contienen diversos algoritmos para el análisis de datos específicos. A continuación, se detallan los tipos de aprendizaje supervisado mencionados [10].

1.4.5.2.1. Algoritmos de regresión

El objetivo de estos algoritmos es obtener un modelo adecuado, ya sea con los datos mismos o con un conjunto de entrenamiento; la salida esperada es un número. Es decir, que no se lo ubica dentro de ningún grupo, sino que devuelve un valor especifico. Estos algoritmos son usados en conjuntos de datos con clase en los cuales se quiere predecir un valor como, por ejemplo; el costo estimado de una vivienda o el valor de temperatura adecuado para el crecimiento de una flor dentro de un invernadero [11] [10].

1.4.5.2.2. Algoritmos de clasificación

Su objetivo obtener el grupo al cual pertenece el elemento de estudio. Este algoritmo clasifica en grupos a ciertos elementos según la similitud de sus características o de un patrón en específico que siguen los datos. Compara los nuevos datos y los ubica en el grupo al cual pertenecen según su similitud, lo que permite predecir que dato estamos tratando. Este tipo de algoritmos son usados en conjuntos de datos con clase de los cuales nos interesa verificar a que conjunto en específico pertenecen. Es muy útil dado que permite realizar ciertas consideraciones en algunos entornos como por ejemplo saber los gustos que tiene un usuario al momento de navegar por internet o por su red social [11] [10].

1.4.5.3. Métricas de evaluación

Su objetivo es estimar que el modelo de clasificación pueda ser usado de manera general para datos futuros y como estos se acoplarían al mismo. En pocas palabras, evalúan el rendimiento del modelo de aprendizaje automático. Existen diversas métricas de evaluación en lo que se refiere a modelos de aprendizaje automático, entre estas se tiene:

- **Matriz de confusión:** Es una tabla en la cual se observa fácilmente en donde el modelo confunde dos clases [12]. Aquí se muestra las siguientes categorías:
    - Verdaderos positivos (TP): cuando la clase real es 1 y se predice un 1
    - Verdaderos negativos (TN): cuando la clase real es 0 y se predice un 0
    - Falsos positivos (FP): cuando la clase real es 0 y se predice un 1
    - Falsos negativos (FN): cuando la clase real es 1 y se predice un 0
- **Exactitud (Accuracy):** Representa el porcentaje de datos clasificados correctamente [12].
- **Sensibilidad (Recall):** Es la tasa de verdaderos positivos. Es decir, el número de datos clasificados como positivos del total de verdaderos positivos [12].

- **Precisión:** Representa al porcentaje de datos clasificados correctamente como positivos de un total de datos identificados como positivos [12].
- **F1-Score:** Promedio entre la precisión y el recall el cual ayuda a verificar si el modelo es perfecto para la clasificación de datos [13]. Su mejor valor es 1.

1.4.6. Tecnologías y herramientas utilizadas en el proyecto

En esta sección se dará a conocer las distintas tecnologías y herramientas útiles para la elaboración del proyecto. El desarrollo de ciertas metodologías se la elaborará haciendo uso del lenguaje de programación Python haciendo uso de las distintas librerías que este entorno de programación nos favorece.

A continuación, se presenta las tecnologías usadas para este proyecto.

1.4.6.1. NLTK

El kit de herramientas de lenguaje natural es un conjunto de módulos de programación, conjunto de datos, tutoriales y ejercicios que cubren el procesamiento de lenguaje natural simbólico y estadístico. NLTK se desarrolló en Python y se distribuye bajo la licencia de código abierto GLP. En el entorno de Python NLTK es una plataforma líder que permite crear programas para trabajar con datos del lenguaje humano, además de proporcionar una interfaz fácil de usar por lo que es considerado una de las herramientas maravillosas para enseñar y trabajar en lingüística computacional y jugar con el lenguaje natural [14].

1.4.6.2. Scikit Learn

Conocida como la biblioteca de aprendizaje automático debido a que permite la integración fácil y rápida de los métodos de aprendizaje automático en el código de Python [15]. Esta biblioteca comprende una amplia gama de métodos de clasificación, regresión, estimación de la matriz de covarianza, reducción de la dimensionalidad, preprocesamiento de datos y generación de problemas de referencia. Está disponible para varios sistemas operativos y su instalación es muy sencilla, se puede acceder a esta biblioteca a través de la URL http://scikit-learn.org.

1.4.6.3. Google Colaboratory

“Colaboratory”, o "Colab” es un producto de Google Research que permite a cualquier usuario escribir y ejecutar código arbitrario de Python en el navegador. Esta herramienta es adecuada para tareas de aprendizaje automático, análisis de datos y educación” [16]. No requiere instalación ni configuración para su uso, además de ser gratuito y ofrecer otros recursos informáticos como GPUs.

2 METODOLOGÍA

Dentro de esta sección se describe el proceso que se lleva a cabo para el cumplimiento de los objetivos del presente trabajo. Su desarrollo consta de dos fases muy importante que son: 1) generación del dataset y 2) desarrollo de un modelo adecuado que permita clasificar las prácticas de transferencia de datos dentro de políticas de privacidad en español.

2.1. Generación de dataset

El proceso de generación del dataset se basa en seis pasos fundamentales que nos ayudarán a recopilar la información necesaria que posteriormente será utilizada para la realización del modelo de clasificación. En esta sección se describe como se adaptaron cada uno de estos pasos para la obtención del dataset adecuado.

2.1.1. Definición de objetivo y alcance

En este paso es importante tener claro en que se va a trabajar y el resultado final de todo el proceso, por lo que debemos limitar nuestro enfoque para no generalizar la información que deseamos recolectar.

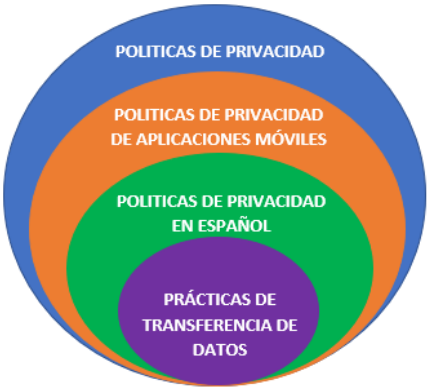


Figura 2.1. Delimitación de objetivo.

De la Figura 2.1 podemos deducir el objetivo y alcance para la generación de nuestro conjunto de datos, que busca etiquetar las prácticas de trasferencia de datos en políticas de privacidad en español de aplicaciones móviles más populares.

2.1.2 Estrategia de selección del conjunto de datos

Actualmente existen diversas políticas de privacidad que permiten la protección de los datos personales, elaboradas por equipos legales y que se actualizan con mucha frecuencia. En nuestro proyecto es importante contar con la mayor diversidad de políticas de privacidad para generar un modelo de clasificación adecuado. Para reflejar esto se toma en cuenta la selección de políticas de privacidad de las aplicaciones móviles más populares, que varían según los siguientes parámetros: (i) el modelo de negocio/servicio de la organización, (ii)

complejidad, determinada por la longitud de las políticas de privacidad, (iii) idioma y (iv) cobertura de los servicios pertenecientes a la organización.

Con base en estos parámetros, se establecen ciertos criterios de clasificación por categoría que se muestran a continuación:

1. Según el modelo de negocio/servicio de la organización
    - Se trabaja con políticas de privacidad de aplicaciones móviles de mayor tendencia.
    - Se limita a políticas que tienen base en Ecuador o que brindan servicios a ciudadanos ecuatorianos.
2. Según su complejidad
    - Se incluyen políticas de distinta complejidad determinada por el tamaño de las mismas.
3. Según el idioma
    - Las políticas de privacidad son en español
4. Según la cobertura de servicios
    - Políticas de privacidad de aplicaciones móviles Android.

En este contexto se establecen también ciertos criterios de exclusión que son:

- Se excluirán aquellas políticas de privacidad que no tengan base en Ecuador.
- Se excluirán aquellos documentos que declaren términos del servicio.

En este punto se verifica el número de descargas de aplicaciones móviles comúnmente usadas (p. ej. Facebook, Instagram, Tiktok). En segundo lugar, se toma una muestra de las aplicaciones con mayor relevancia y las organizamos por sectores (p.ej. Artes, Compras, Redes sociales, Noticias). En Anexo I se aprecia las políticas de privacidad seleccionadas para el presente trabajo de integración curricular.

2.1.3 Definición del tamaño de la muestra

El número de políticas de privacidad que nos ayudará a desarrollar el modelo de clasificación se lo establece según ciertos razonamientos como: (i) razonamiento por analogía, (ii) razonamiento por experticia, y (iii) el uso de heurísticas estadísticas.

En el presente proyecto se hace uso del razonamiento por analogía dado que se inicia nuestro estudio con el número de muestras con las que se obtuvo un buen performance en otros proyectos. Aproximadamente un 87% de rendimiento se obtuvo con 100 políticas de privacidad por lo que nuestro dataset está conformado por el mismo número de políticas. Con esto se espera tener un rendimiento mayor o igual, luego de varias pruebas y sus respectivos análisis con diferentes algoritmos de clasificación.

2.1.4 Definición del esquema de etiquetado

Aquí se define el protocolo utilizado para el etiquetado de los datos extraídos de las distintas políticas de privacidad.

Según la guía de transparencia relacionada con la protección de datos personales bajo el Reglamento General de Protección de Datos se estipula que: “la transparencia es una obligación general… en relación con el procesamiento de datos” [1]. Por tal motivo menciona que con las políticas de privacidad debe quedar claro el tratamiento que se le dará a los datos de usuario y que estas deben ser de fácil acceso dentro de las aplicaciones que las contengan.

Lo mismo dice la Ley Orgánica de Protección de Datos Personales; específicamente hablando de la transferencia de datos menciona que los datos pueden compartirse siempre y cuando se cumplan las funciones legitimas establecidas en la Ley. En este contexto, es indispensable conocer ciertos parámetros que permitan identificar a esta práctica de privacidad, y así determinar qué datos se comparte o tiene acceso otra entidad (terceros).

En proyectos elaborados en relación con el etiquetado de tratamiento de datos personales se ha establecido ciertos esquemas para la anotación de diferentes prácticas de tratamiento de datos, dentro de los cuales se establecen ciertos parámetros para su identificación y clasificación. Uno de ellos es el proyecto “UsablePrivacy” [17] el cual es una página web que permite verificar prácticas de privacidad dentro de políticas de privacidad en inglés de aplicaciones y páginas más conocidas internacionalmente.

Con base en estos dos proyectos establecemos nuestro esquema de etiquetado para la práctica de transferencia de datos. En la Tabla 2.1 se presentan los elementos de transparencia propuestos para nuestro esquema.

Tabla 2.1. Elementos de transparencia de esquema de etiquetado.

Elementos de transparencia	**Descripción**
Comparte/No Comparte	Útil para identificar que la política de privacidad establece si se comparte o no algún dato.
Tipo de información personal	Útil para verificar que tipo (categoría) de información personal podría ser compartida por la aplicación con terceras entidades.
Propósito	Indica el objetivo de una tercera entidad para recibir o recolectar la información personal del usuario.

Tercera entidad	Identifica a la tercera entidad involucrada en la compartición de los datos.
Tipo de recolección	Identifica el mecanismo usado por la tercera entidad para obtener o recolectar la información personal del usuario.
Identificable	Indica si en la política de privacidad se establece que los datos podrían identificar al usuario o no.
Consentimiento/ Base legítima	Indica la selección de información personal que explícitamente es compartida por el usuario.

Debido a las diversas categorías de información que existen dentro de los elementos de transparencia establecidos, se determinan ciertos tipos para identificar de mejor manera la práctica de privacidad; es así que se define lo siguiente:

1. Comparte/No comparte

   0 = No comparte

   En este caso el tipo de dato dentro de este elemento de transparencia identifica que la aplicación no comparte información con una tercera entidad.

   1 = Comparte

   Este tipo de dato identifica que la aplicación comparte cierta información con una tercera entidad.

2. Tipo de información

   1 = Información personal

   El tipo de dato identifica que la categoría de información que es compartida por la aplicación es personal (ej. Fotos, videos, contactos, notas de voz, etc.)

   2 = Perfil de usuario

   En este caso la aplicación comparte información que se encuentra dentro del perfil de usuario como: foto de perfil, likes, comentarios, preferencias, etc.

   3 = Información de contacto

   La información de contacto compartida por la aplicación sería, por ejemplo: nombre, número telefónico, email, código postal, etc.

   4 = Identificadores del dispositivo

Este tipo de dato identificaría que la aplicación comparte información del terminal del equipo en el cual se está usando la app, esta podría compartir: IMEI, MAC address, SSID, Serial number, etc. Según la GDPR, estos datos no están enlazados a datos de contacto del usuario, pero si pueden servir para aprender algo del mismo [3].

5 = Actividad

En este caso la actividad a la cual nos referimos es el historial de uso de la app por parte del usuario, además podría ser el historial de llamadas, el historial de chats, etc.

6 = Localización/Dirección

Tipo de información delicada pero también muy útil en ciertos casos. En este caso también se podría mencionar:

Localización: La aplicación comparte la ubicación dinámica del usuario, por ejemplo, la ubicación en tiempo real de WhatsApp.

Dirección: Este tipo de información hace distinción del anterior dado que la aplicación aquí comparte una ubicación fija, por ejemplo, dirección postal, dirección fija del GPS, dirección IP, etc.

7 = Información financiera

En ocasiones ciertas aplicaciones brindan ciertos servicios con algunas cuotas, por tal motivo la aplicación podría compartir información bancaria, tarjetas de crédito, etc.

8 = Anónima

En ocasiones las aplicaciones comparten información anónima a otras entidades con la finalidad de tener algún servicio.

3. Propósito

El propósito de la compartición de datos es importante ya que esto se estipula en la LOPDP, si el objetivo de la compartición no está dentro de las normas legales la aplicación podría identificarse como peligrosa.

1 = Operación y seguridad del servicio

En ocasiones se puede requerir datos para el funcionamiento adecuado de la app, es por eso que se puede compartir información para: mantener una sesión abierta, para autenticación, detección de bugs, etc.

2 = Análisis e investigación

En ciertos casos es importante la compartición de datos para realizar un estudio, ya sea para el funcionamiento de la app o para estudios epidemiológicos de interés público tal como se estipula en la LOPDP.

3 = Publicidad

Los datos compartidos pueden ser usados para generar publicidad según las preferencias del usuario, según lo dicte el consentimiento del propietario de los datos.

4 = Requisito legal

La LOPDP establece que en ocasiones los datos pueden ser compartidos sin el consentimiento del propietario de los datos cuando tengan que proporcionarse a autoridades administrativas o judiciales.

5 = Servicio básico

Los datos proporcionados por la aplicación pueden verificar la edad y otras características necesarias para la funcionalidad prevista de la aplicación.

6 = Servicio Adicional

Los datos son importantes para brindar ciertos servicios técnicos a la aplicación

7 = Personalización

Los datos compartidos pueden ser necesarios para personalizar el contenido de la aplicación, tal como sucede con la publicidad (ej. Lenguaje preferido)

8 = Traspaso de dominio

En ocasiones los datos personales se comparten con otra organización u dominio cuando se realiza una compra o venta de los derechos de una organización a otra.

9 = Mejorar contenido

Al igual que para personalizar el contenido que se pueden ver en distintas plataformas, los datos se pueden compartir para mejorar el contenido que tiene una aplicación luego de realizar ciertos estudios.

4. Tercera entidad

1 = Otra parte de la organización

La cobertura de servicios de una organización va más allá de sí misma. En ocasiones se tiene diferentes servicios dentro de una sola organización, por lo que la información podría ser compartida entre los servicios de una misma entidad.

2 = Tercero designado

La información es compartida y visible para el administrador de la aplicación, así como para aquellos que brindan soporte a los usuarios dentro de la organización.

3 = Tercero no designado

Es posible compartir información personal con otras entidades solo si se tiene el consentimiento del propietario de los datos tal como lo estipula la LOPDP.

4 = Público

En LOPDP se menciona que se pueden compartir datos siempre y cuando estos no permitan la identificación del usuario, en pocas palabras que sean datos anonimizados.

5. Tipo de recolección

   1 = Recibe directamente de la aplicación

   Relacionado a que terceras personas reciben información a través de la aplicación.

   2 = Rastreo de la aplicación

   Puede usar técnicas similares a las cookies que son usadas en páginas web.

   3 = Otros

6. Identificable

   0 = No permite identificar al usuario

   En la LOPDP se establece que cierta información puede ser compartida al público con esta no permite la identificación del usuario.

   1 = Permite identificar al usuario

   Según la LOPDP solo se puede compartir información que identifique al usuario si la organización tiene el consentimiento del propietario de los datos.

7. Consentimiento/base legitima

   1 = Consentimiento

Este campo indica diversos factores relacionados al consentimiento del usuario en la compartición de sus datos personales. Se puede referir a:

Opt-out: En este caso el usuario da el consentimiento de compartir cierta información, aquí el usuario debe seleccionar la información que no desea compartir.

Opt-in: En este caso el usuario da el consentimiento de compartir cierta información identificable. El usuario debe seleccionar la información personal a compartir.

2 = Obligación legal

El campo denota una compartición de datos personales mediante una obligación legal, bajo las leyes establecidas en la LOPDP. No necesita el consentimiento del propietario de los datos para su compartición.

3 = Interés público

Los datos serán usados para verificar necesidades sociales, no se pide el consentimiento del propietario por lo que en su uso debe primar criterios equitativos y de bien común.

4 = Obligación contractual

El valor indica que la compartición de datos es indispensable para el requerimiento o adquisición de un servicio (p.ej. Compras en línea, inscripción a cursos educativos).

5 = Interés vital

Este valor indica que terceras entidades acceden a datos personales sin el consentimiento del propietario debido a temas de salud. (p.ej. Estudios de propagación de contagios de covid-19)

6 = Acceso público

Denota que los datos serán parte de una base de datos de acceso público para la elaboración de estudios.

7 = Interés legítimo

En este caso en particular se debe especificar la frase “interés legítimo” para marcar el campo con dicho valor.

Los parámetros y códigos establecidos anteriormente son escogidos según los reglamentos establecidos en la LOPDP relacionados a la transferencia de datos personales a terceros. Dentro de los cuales se establece la compartición de datos por parte del encargado de los datos, por parte de terceros y las excepciones que se tiene en la transferencia de datos según el consentimiento del usuario. Se trata de abarcar todo los del reglamento para identificar la práctica de transferencia de datos en las distintas políticas de privacidad en español de aplicaciones Android más populares.

2.1.5 Definición de procedimiento de etiquetado

El esquema de etiquetado propuesto anteriormente conlleva una serie de validaciones que permitirá el etiquetado de una política de privacidad. Este proceso se lo define de mejor manera dentro de un diagrama de flujo el cual se presenta en el Anexo II en el cual se describen ciertas condiciones que deben cumplirse para etiquetar que en una política de privacidad existe una práctica de transferencia de datos.

2.1.5.1 Validación de procedimiento de etiquetado

Para validar el proceso descrito en el diagrama de flujo, se realizó el etiquetado de 15 políticas de privacidad de manera manual siguiendo dicho proceso.

En primera instancia se toma una política de privacidad, se la divide en segmentos y según una codificación de colores se extrae subsegmentos donde se comprueba cada una de las validaciones del diagrama de flujo de manera general. Ejemplo de esto se lo puede ver en la Figura 2.2:

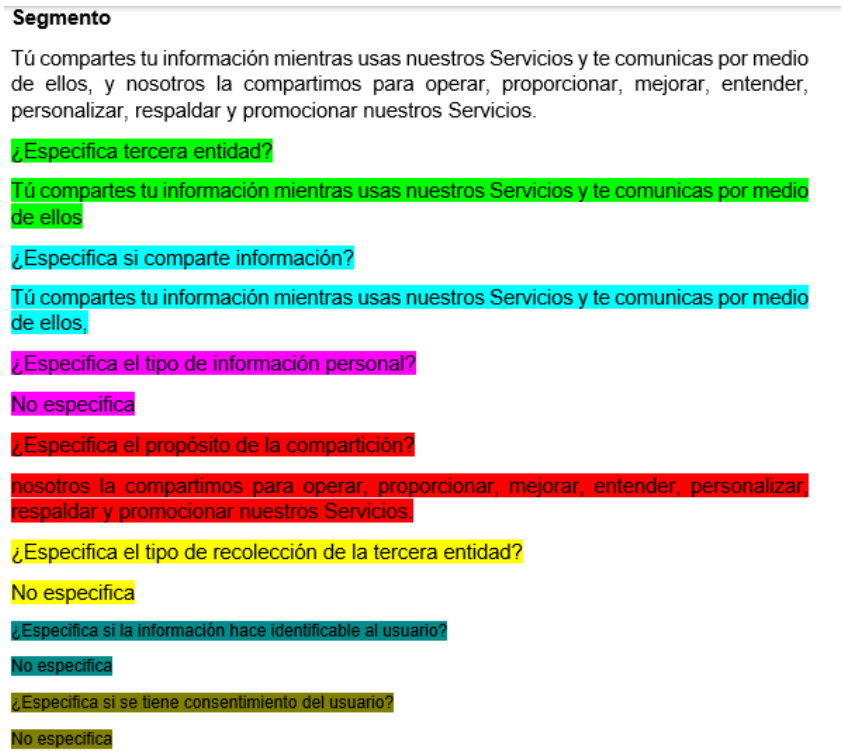


Figura 2.2. Ejemplo de etiquetado manual.

La explicación a lo visto en la Figura 2.2 es la siguiente:

Verde: Con este color se denota la validación principal para verificar si este segmento trata de una práctica de transferencia de datos, en este caso se verifica si el segmento habla de alguna tercera entidad.

Celeste: Aquí se denota si el segmento habla sobre la compartición de datos o de alguna transferencia de información.

Morado: Denota si el segmento especifica algún tipo de información personal del usuario.

Rojo: Denota si dentro del segmento se especifica el propósito por el cual se comparte información del usuario.

Amarillo: Denota si en el segmento se especifica el tipo de recolección de la tercera entidad.

Verde azulado: Denota si la información que contiene el segmento hace identificable al usuario.

Oliva: Denota si en el segmento se especifica de manera clara que se tiene el consentimiento de usuario para su compartición.

Durante este proceso se encuentran una serie de dificultades, entre ellas:

- La división por segmentos de gran magnitud en ocasiones no era tan satisfactoria dado que en algunos de ellos no se encontraba la información necesaria para validar alguna de las condiciones establecidas en el proceso.
- Al verificar las condiciones establecidas en el diagrama de flujo junto con el esquema de etiquetado se observa que algunos elementos de transparencia son ambiguos.
- La condición principal no basta para identificar que el segmento hable de una transferencia de datos, dado que solo valida si se menciona una tercera entidad.

Al analizar estos problemas se da una solución a cada uno:

- Se establece que el análisis de etiquetado se hará por segmentos de menor tamaño.
- Se crea subtipos dentro de los elementos de transparencia necesarios para etiquetar de mejor manera las políticas de privacidad.

- Se establecen dos condiciones principales para identificar rápidamente que en el segmento se habla de una práctica de transferencia de datos.

Luego de hacer las respectivas correcciones se tiene como resultado el diagrama de flujo presente en el Anexo III que describe el procedimiento de etiquetado de manera actualizada.

2.1.5.2 Uso de esquema y procedimiento de etiquetado

Una vez modificado el procedimiento y esquema de etiquetado, es necesario ponerlos en práctica para corroborar su uso dentro de la herramienta de anotación.

En este punto con ayuda de las hojas de cálculo se genera una tabla con todos los valores y elementos de transparencia del esquema de etiquetado. Se analiza cada uno de los segmentos de las políticas de privacidad y se apunta el valor indicado según el contexto de los mismos.

En la Figura 2.3 se presenta un ejemplo que muestra la codificación de los segmentos de las políticas de privacidad según el esquema de etiquetado.

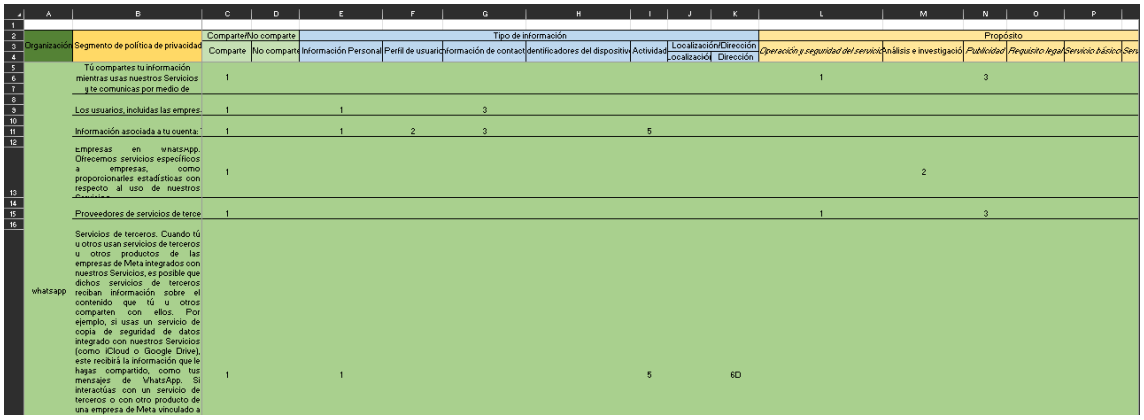


Figura 2.3. Uso de codificación para anotación de políticas de privacidad.

Se observa que con este modelo se puede validar si dentro de la política de privacidad estudiada se habla de prácticas de transferencia de datos.

2.1.5.3 Uso de herramienta de etiquetado

La Figura 2.4 muestra de manera resumida la funcionalidad de la herramienta de anotación desarrollada por Félix Miño en su trabajo de titulación [3].

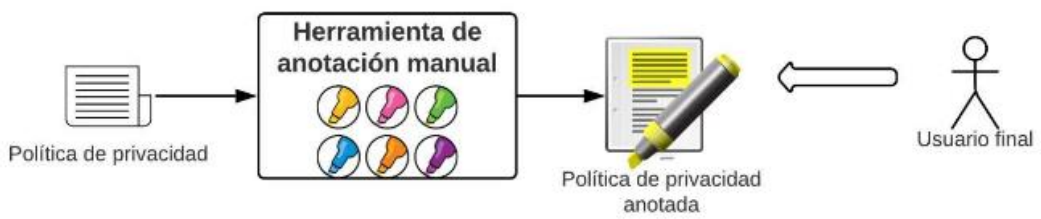


Figura 2.4. Resumen funcionalidad de herramienta de anotación **[3]**.

2.1.5.4 Etiquetado de políticas de privacidad con herramienta de anotación

Primero se debe cargar las políticas de privacidad en formato txt dentro de la herramienta de anotación y se coloca los valores, atributos y tratamientos que serán parte del etiquetado. En este caso, cargamos la práctica de transferencia de datos, sus elementos de trasparecía y los tipos específicos de cada elemento, que fueron definidos anteriormente en nuestro esquema de etiquetado. Procedemos a realizar nuestro etiquetado de políticas de privacidad para esto se selecciona el texto que esté relacionado con la práctica de transferencia de datos y se coloca la etiqueta respectiva del lado derecho, un ejemplo de esto se muestra en la Figura 2.5.

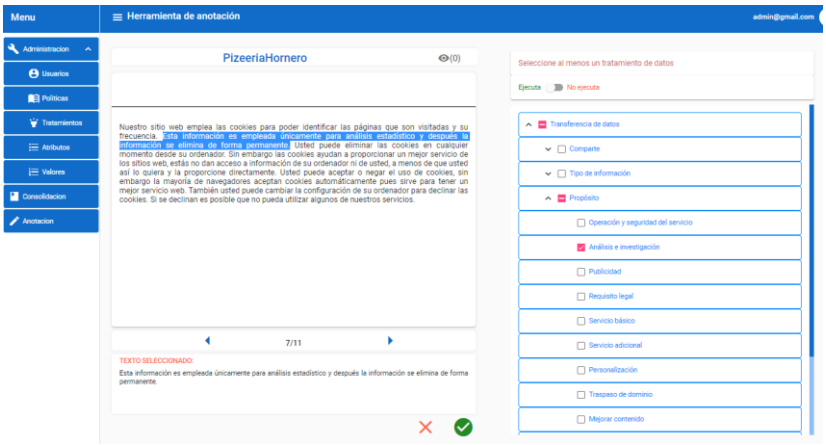


Figura 2.5. Interfaz para anotación de políticas de privacidad.

Al acabar la anotación basta con hacer clic en el visto de color verde para guardarla. Las anotaciones se guardan en nuestra base de datos y con ayuda de una consulta para obtener la información necesaria se la extrae en un archivo CSV. En el Enlace 1 del Anexo VI se puede apreciar el archivo obtenido. Posteriormente este se lo pasa a archivos en formato YAML que nos ayudarán a formar el dataset correspondiente para nuestro análisis. Los archivos los podemos apreciar en el Enlace 2 del Anexo VI.

2.1.5.5 Generación de dataset en Python

Este proceso tiene como objetivo pasar los archivos yaml a un dataframe con ayuda de algunas librerías y funciones de Python. La conversión inicia leyendo todos los archivos YAML y con ayuda del método “chdir” de la librería OS se facilita el manejo y lectura de los mismos.

Este proceso se lo hace dentro de un ciclo for y se utiliza algunas funciones de las librerías yaml y pandas para la construcción del conjunto de datos. La Figura 2.6 muestra el código utilizado para la lectura de archivos y la construcción del dataframe.

```
#Main
for file in os.listdir():
    if file.endswith(".yaml"):
        file_path = f"{path}/{file}"
        with open(file_path, 'r') as d:
            data = yaml.safe_load(d)
        daux = pd.json_normalize(data["segments"], "annotations", ["segment_text"])
        df = pd.concat([df,daux])
```

Figura 2.6. Lectura de archivos YML para elaboración de dataset.

En ocasiones los datos vienen con un índice por defecto por lo que es importante quitarlo. Además, de eliminar palabras comunes que no brindan significado dentro del análisis de datos, quitar los signos de puntuación y convertir cada palabra a su raíz. Esto hará más fácil la identificación de las palabras durante el proceso. La Figura 2.7 muestra el dataset resultante luego de aplicar los filtros adecuados.

	segmento_texto	anotaciones
0	polit privac compañ comprend dentr adid group ...	[{'practica': 'TRANSFERENCIA&Otra parte de la ...
1	¿qu tip inform recopil exist cas adid podr ped...	[{'practica': 'TRANSFERENCIA&Información perso...
2	utiliz 'cookies' reun inform visit hag siti we...	[]
3	usted visit siti web reun dat conoc 'fluj clic...	[]
4	uso inform visit registr utiliz proteg tod inf...	[]
...	...	...
61	marketing promocion public tercer permit zoom ...	[]
62	autent integr segur proteccion autentic cuent ...	[]
63	comunic usted utiliz dat personal inclu inform...	[]
64	razon legal cumpl legisl aplic respond proces ...	[]
65	proteg interes fundamental tercerostrat determ...	[]

Figura 2.7. Dataset resultante.

Posteriormente se puede aplicar ciertos preprocesamientos de datos que permitirán adecuar más el conjunto de datos y así aplicar algoritmos de análisis de Machine Learning.

2.2. Desarrollo de modelo de clasificación

Su desarrollo contiene siete pasos fundamentales que nos ayudan de cierta manera a: limpiar la información, procesar los datos de mejor manera, configurar nuestros algoritmos de clasificación y evaluarlos para conseguir un modelo adecuado que permita etiquetar las

prácticas de transferencia de datos. Estos pasos se los explica más a detalle en las siguientes secciones.

2.2.1 Preprocesamiento de datos

En la sección anterior se evidencia que los datos obtenidos no contienen un formato adecuado para analizarlos bajo mecanismos ML. El vectorizar dicha información hace más sencillo el uso de esto algoritmos. La librería sklearn y su módulo feature_extraction.text del cual importamos las clases CountVectorizer y TfidfVectorizer nos facilitará este proceso.

Estos métodos permiten crear vectores con palabras del texto ingresado, dándoles un cierto valor para ser procesados. El objetivo de usar estas librerías es crear una matriz dispersa con las palabras esenciales de cada segmento y así tener una representación numérica de las políticas de privacidad. Luego de la preparación de los datos, el dataset resultante quedaría tal como se muestra en la Figura 2.8.

colectivamente	respetan	...	integren	phone	reunión	fusiones	adquisiciones	reestructuraciones	afecten	tratarse	externossi	Clase
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&Otra parte de la organización
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&Información personal&Información...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Perfil de usuario&Ter...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&Operación y seguridad del servic...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&Información personal&Operación y...
...	...	...	...	...	...	...	...	...	...	...	...	...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Información personal&...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Información personal&...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Información personal&...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Información personal&...
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	TRANSFERENCIA&SiComparte&Información personal&...

Figura 2.8. Vectorización de dataset.

La última columna representa la etiqueta de salida que será de gran ayuda para el desarrollo de nuestro modelo de clasificación.

2.2.1.1 Variable categórica a numérica

Contar con variables binarias es de gran importancia al momento de desarrollar un modelo de clasificación basado en el procesamiento de texto. Los diferentes algoritmos de Machine Learning demandan que exista variables binarias que denoten la clasificación de cada uno de los datos. Al realizar dicho cambio el dataset resultante se vería como el de la Figura 2.9

r	adid	group	denomin	colect	adidas	...	pen	perjuri	integrarlosest	proteccionsi	zoomiq	comercialic	acusatori	tercerostrat	TransferenciaDatos
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	1
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	1
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.430997	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.222860	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.287152	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.343590	0.0	0.0	0.0	0.0	0.0	0
.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0

Figura 2.9. Dataset con variable binaria.

2.2.2 División de dataset

Dentro del aprendizaje automático el dividir un dataset en conjunto de entrenamiento y prueba es de gran importancia. Esta división permite tener dos conjuntos de datos que permiten entrenar el modelo y luego evaluar su rendimiento. Definir las variables X, Y nos facilitará este proceso, con ayuda de la clase train_test_split de la librería sklearn.model_selection creamos nuestros conjuntos de prueba y entrenamiento. La Figura 2.10 muestra el uso de esta librería en el entorno de desarrollo de Python.

```
##Dividir el dataset
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.3,random_state=4)
print("Conjunto de entrenamiento: ",xtrain.shape,ytrain.shape)
print("Conjunto de validación: ",xtest.shape,ytest.shape)
```

Figura 2.10. Código para división en conjunto de entrenamiento y prueba.

De esta imagen podemos mencionar que:

- Los atributos “X” y “Y” son las variables que contienen las características y etiqueta salida del conjunto de datos respectivamente.
- Test_size define el tamaño que tendrá el subconjunto de prueba, en este caso el tamaño será un 30% del tamaño del dataset original. Por consiguiente, el 70% restante será para el conjunto de entrenamiento.
- Random_state permite controlar el generador de números aleatorios.

Estos conjuntos serán útiles para realizar las configuraciones adecuadas de nuestro algoritmo ML.

2.2.3 Configuración de algoritmo

El rendimiento del modelo a desarrollar depende en gran cantidad de la configuración que se les dé a los distintos algoritmos de clasificación. Al hablar de configuración nos referimos a

verificar los parámetros adecuados de las distintas funciones que nos brinda el lenguaje Python, y que nos permitan utilizar los distintos clasificadores de ML. Los siguientes algoritmos supervisados fueron usados en el presente proyecto:

- SVM: Su objetivo es crear un límite de decisión, más conocido como hiperplano, el cual ayuda a clasificar un dato en la categoría correcta. Este hace uso de puntos o vectores extremos que sirven de soporte para la creación del límite [18].
- Regresión logística: Su clasificación se basa en organizar cada uno de los datos analizados en valores probabilísticos entre 0 y 1 como si fuera un tipo de dato booleano [19].
- Árbol de decisiones: Su división se basa en preguntas de si o no y según los resultados en cada nivel arma la estructura del árbol hasta obtener las predicciones necesarias [20].
- Naive Bayes: Se basa en el teorema probabilístico de Bayes y es muy utilizado en problemas de clasificación de texto. Utiliza la probabilidad para hacer sus predicciones por lo que es útil para la construcción de modelos que requieran un aprendizaje automático rápido [21].
- KNN: Este algoritmo agrupa un dato a la categoría del dato o datos más cercano/s y con el que tenga mayor similitud. El número de vecinos cercanos lo podemos definir nosotros dentro de la función [22].

```
def classifier(clf):
  if clf == "SVM":
    return svm.SVC(kernel="rbf",class_weight="balanced")
  if clf == "RegresionLogistica":
    return LogisticRegression(C=1.0,penalty='l2',random_state=1,solver="newton-cg",class_weight="balanced")
  if clf == "KNN":
    return KNeighborsClassifier(n_neighbors=4)
  if clf == "NaiveBayes":
    return MultinomialNB()
  if clf == "DecisionTree":
    return DecisionTreeClassifier(criterion='entropy', max_depth=2, random_state=0, class_weight='balanced')
```

Figura 2.11. Configuración de algoritmos de ML.

La función kernel o núcleo presente en el clasificador SVM de la Figura 2.11 se usa comúnmente para pasar de un espacio de dimensiones pequeñas a otro de mayores dimensiones y así cuantificar de mejor manera la similitud entre dos observaciones [23]. En nuestro proyecto hacemos uso de la función de base radial (rbf) dado que permite asociar datos continuos a un conjunto de datos discretos [24].

Con esto se tiene todos los algoritmos configurados para dar paso al entrenamiento del mismo. Cabe mencionar también que los parámetros utilizados no son los únicos y que pueden variar según las necesidades del caso.

2.2.4 Entrenamiento de dataset

El uso del conjunto de entrenamiento es sustancial para realizar el ajuste o entrenamiento de los distintos modelos de clasificación que luego serán puesto a prueba con ayuda del conjunto restante (prueba).

Se le proporciona como variables de entrada a cada clasificador aquellas variables resultantes de la división del dataset original, siendo “Y” la variable que contiene la clase de salida y “X” las demás características del dataset. En la Figura 2.12 se presenta un ejemplo del ajuste del modelo de clasificación.

```
#Ajuste o entrenamiento
c_svm.fit(xtrain,ytrain)
```

Figura 2.12. Entrenamiento para modelo SVM en Python.

El entrenamiento es similar para los demás algoritmos y se usa las variables X, Y de entrenamiento. Estos modelos ayudarán a realizar las predicciones necesarias para posteriormente evaluar su rendimiento.

2.2.5 Predicción de modelo de clasificación

El cálculo de predicciones o hipótesis es de gran importancia para la evaluación del modelo entrenado. Para predecir el resultado del conjunto de pruebas se hace uso de la función “predict” definida en cada algoritmo. A esta se le pasa como variable de entrada el conjunto de prueba que contiene las características de cada política de privacidad, en este caso los segmentos de texto. Su sintaxis es muy sencilla y se presenta un ejemplo en la Figura 2.13.

```
#hipotesis o prediccion de salidas
h=c_svm.predict(xtest)
```

Figura 2.13. Cálculo de predicciones para modelo SVM.

Con las predicciones obtenidas de cada modelo podemos evaluar cada uno de ellos y así verificar el modelo adecuado para la clasificación de prácticas de transferencia de datos.

2.2.6 Evaluación de modelos de clasificación

Las métricas de evaluación de cada modelo explican el rendimiento que tiene el modelo al momento de clasificar los datos ingresados en ellos. Aquí se hace una comparación de datos del conjunto de prueba (datos reales) con los valores obtenidos en la predicción del modelo. Las diferentes métricas de evaluación facilitan la observación del rendimiento del modelo, metricas como la precisión, el error cuadrático medio, la matriz de confusión, etc., son las más utilizadas dentro de los modelos de clasificación.

Las funciones que nos proporciona Python serían útiles para el cálculo de dichos valores. Su resultado indicará que modelo es el adecuado para la clasificación de prácticas de transferencia de datos.

2.2.7 Extracción del modelo de clasificación

La exportación del modelo no es más que extraer dicho modelo entrenado en un archivo de extensión pkl para luego ser utilizado en la clasificación de otros conjuntos de datos, así como armar un ensamble de clasificadores. Hacer esto en Python en muy sencillo y con el uso de la librería “pickle” que junto con su método dump permite guardar el modelo con dicha extensión. La Figura 2.14 muestra el uso de esta función para la extracción del mejor modelo.

```
filename = '{}_model.pkl'.format(outfile)
pickle.dump(text_clf, open(filename, 'wb'))
```

Figura 2.14. Código para exportación de modelo en Python.

2.2.8 Selección de modelo apropiado

Con el proceso descrito en las secciones 2.2.1 hasta la sección 2.2.7 podemos realizar varias pruebas con distintos parámetros y algoritmos para verificar el modelo adecuado que clasificará cada una de las etiquetas escogidas como salida. Se escogen aquellos valores con mayor número de anotaciones para elaborar cada modelo clasificador.

Se elaboraron seis modelos distintos para: i) que clasifique cuando un segmento tenga una anotación de transferencia (general), ii) que clasifique los segmentos en los cuales se diga si se comparte información, iii) que clasifique los segmentos donde se mencione que se comparte información personal, iv) que clasifique segmentos donde se mencione que la información se comparte con un tercero designado, v) que clasifique segmentos donde se mencione que la información que se comparte identifica al usuario y vi) que clasifique segmentos donde se mencione que la información se la recibe directamente de la aplicación.

2.2.9 Ensamble de clasificadores

El ensamble o combinación de modelos tiene como objetivo principal el de mejorar la precisión de las predicciones [25]. Estará conformado de los modelos de mejor rendimiento antes seleccionados. Por lo general la precisión obtenida suele ser mayor que la precisión de cada componente.

La Figura 2.15 muestra el enfoque que se le da al ensamble de clasificadores, el cual estará conformado por dos modelos: i) el clasificador de anotaciones de transferencia de datos y ii) uno de los clasificadores de un valor en específico.

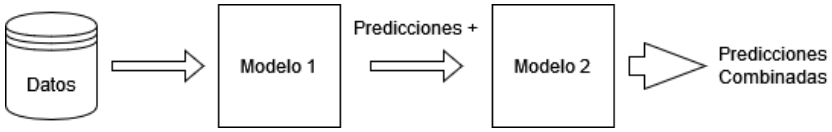


Figura 2.15. Flujo de modelo ensamblado.

Los datos serán ingresados al primer clasificador del cual se tiene ciertas predicciones. De estas se seleccionan únicamente las positivas, que pasarán al siguiente clasificador para verificar si existe el valor especificado. El resultado obtenido serán las predicciones combinadas de los dos modelos las cuales se usarán para calcular las métricas de evaluación.

El conjunto de datos de entrada para validar este clasificador será la unión de los conjuntos de prueba sobrantes del entrenamiento de los modelos individuales. Aquí se determina ciertas etiquetas de salida que se las presenta en la Tabla 2.2.

Tabla 2.2. Etiquetas de salida para conjunto de datos de validación.

Cod.	Etiqueta	Descripción
0	NULL&NULL	Cuando no tiene una anotación de transferencia y no tiene el valor.
1	TRANSFERENCIA&NULL	Cuando tiene una anotación de transferencia, pero no tiene valor
2	TRANSFERENCIA&VALOR	Cuando tiene una anotación de transferencia y valor
3	NULL&VALOR	Cuando no tiene una anotación de transferencia y tiene valor

Aquellos segmentos que se identifiquen con una etiqueta de transferencia de datos pasarán al siguiente clasificador, este verificará si se encuentra el valor especifico y hará sus predicciones. La Tabla 2.3 muestra el enfoque dado a estas etiquetas dentro del ensamble.

Tabla 2.3. Codificación binaria en clasificadores del modelo ensamblado.

Cod.	Cod. Modelo 1	Cod. Modelo 2	Salida
0	0	-	0
1	1	0	0
2	1	1	1
3	0	-	0

Dentro del proceso de ensamble a estás etiquetas se la procesa de manera binaria, entonces:

- En el modelo 1:
    - Aquellos segmentos con etiquetas de código 0 y 3 (primera columna Tabla 2.3) se las codifica como 0 (segunda columna Tabla 2.3).
    - Los segmentos con etiquetas de código 1 y 2 (primera columna Tabla 2.3) se los toma como 1 (segunda columna Tabla 2.3).

Aquellos segmentos de codificación 1 pasarán al siguiente clasificador como una nueva entrada de datos, entonces se tiene lo siguiente:

- En modelo 2:
    - Aquellos segmentos de código 1 (primera columna Tabla 2.3) se los toma como 0 (tercera columna Tabla 2.3).
    - Los segmentos de código 2 (primera columna Tabla 2.3) se los toma como 1 (tercera columna Tabla 2.3).

Finalmente, se tiene las predicciones del modelo. Estas son comparadas con la salida real (última columna de la Tabla 2.3), con el objetivo de calcular las métricas y evaluar el rendimiento obtenido.

El proceso de ensamble elaborado en Python se lo puede apreciar en la Figura 2.16 donde se muestra un ejemplo del proceso descrito.

```
# Definir las variables
x=np.asarray(df['segmento_texto'])
result = model1.predict(x)
#Agrego predicciones a dataset
df['predicccciones_class1']=result
#Quito predicciones de 0
df.drop(df[(df['predicccciones_class1'] == 0)].index, inplace=True)
#Defino nuevas variables x y y
x1 = np.asarray(df['segmento_texto'])
y1 = np.asarray(df['practicaBin'])
#predicciones segundo clasificador
result1 = model2.predict(x1)
```

Figura 2.16. Proceso de ensamble en Python.

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

Los procesos descritos en la sección 2.2 del presente trabajo tienen como objetivo el desarrollo del mejor modelo de aprendizaje automático que permita clasificar la práctica de transferencia de datos. Para esto se toma dos enfoques: i) evaluar el modelo que clasifica

correctamente si se tiene una anotación de transferencia de datos y ii) evaluar el modelo que clasifique los valores dentro de los elementos de transparencia. Las métricas calculadas ayudan a verificar que modelo es el mejor. En esta sección se da a conocer los resultados obtenidos luego de varias pruebas elaboradas; dentro de las cuales también se cambia ciertos parámetros hasta comprobar el mejor rendimiento y cual modelo lo tiene.

3.1.1 Resultados obtenidos con dataset natural

- **Primer enfoque**

Al ingresar un dataset únicamente quitando los signos de puntuación, palabras comunes y transformando cada palabra a su raíz se tiene:

Tabla 3.1. Valores de métricas para clasificador general.

Métricas	SVM	R. Logística	Naive Bayes	KNN	Decision Tree
Exactitud	92.91%	92.12%	71.27%	91.18%	91.43%
Precisión	83.80%	85.71%	18.55%	72.5%	60.09%
Recall	41.12%	30.84%	50.46%	27.10%	57.00%
F1 Score	55.12%	45.36%	27.13%	39.45%	58.51%

La Tabla 3.1 muestra los resultados obtenidos para el accuracy, precisión, recall y f1 score de los distintos clasificadores implementados. Los resultados de recall y f1 score son los más relevantes en el contexto del presente trabajo. Una mejor representación de los datos se tiene en la Figura 3.1 en donde se aprecia que el mejor valor de recall y f1 score es para el clasificador árbol de decisión, lo cual se corrobora también con los datos de la Tabla 3.1.

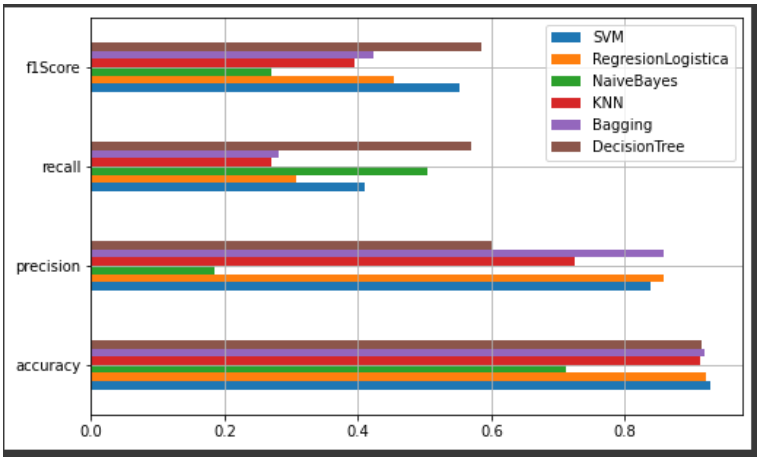


Figura 3.1. Valores de métricas de cada clasificador.

Los resultados obtenidos para este caso no son aceptables dado que un recall por debajo del 70% indica que no se está clasificando correctamente los datos ingresados en la clase

específica. La matriz de confusión de la Figura 3.2 muestra el desbalance que tienen los datos los cual provoca dichos valores de evaluación

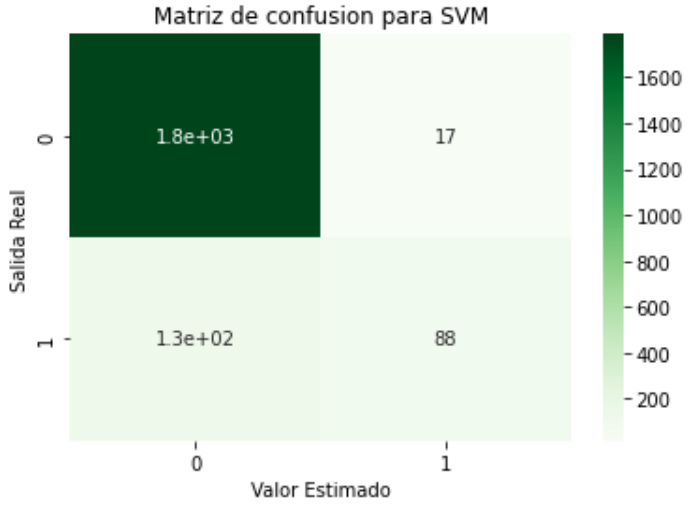


Figura 3.2. Matriz de confusión obtenida con modelo SVM.

Se ve que el clasificador trabaja de mejor manera cuando se detecta aquellos segmentos que no tienen una anotación de transferencia de datos. A simple vista se concluye que se tiene mayor concentración de ellos cuando el valor real es 0 y el predicho también es 0 (color con mayor tonalidad) es decir, que existe mayor cantidad de segmentos que pueden tener una práctica de transparencia, pero que no tiene nada que ver con una transferencia de datos.

- **Segundo enfoque**

Aquí se ingresa el dataset en el cual solo existen aquellos segmentos donde se tiene una anotación de transferencia de datos, de la misma manera sin realizar ningún otro tipo de preprocesamiento. La Tabla 3.2 muestra los valores obtenidos luego procesar los datos mencionados.

Tabla 3.2. Valores de métricas para clasificador de valor "Información personal".

Métricas	SVM	R. Logística	Naive Bayes	KNN	Decision Tree
Exactitud	81.63%	80.40%	70.20%	75.51%	69.79%
Precisión	82.83%	80.99%	84.91%	82.54%	83.33%
Recall	97.47%	98.98%	76.76%	88.38%	78.28%
F1 Score	89.55%	89.09%	80.63%	85.36%	80.72%

Los resultados son alentadores dado que, con un dataset sin mucho procesamiento, se clasifica correctamente el valor especificado, que en este caso es “se comparte información personal”.

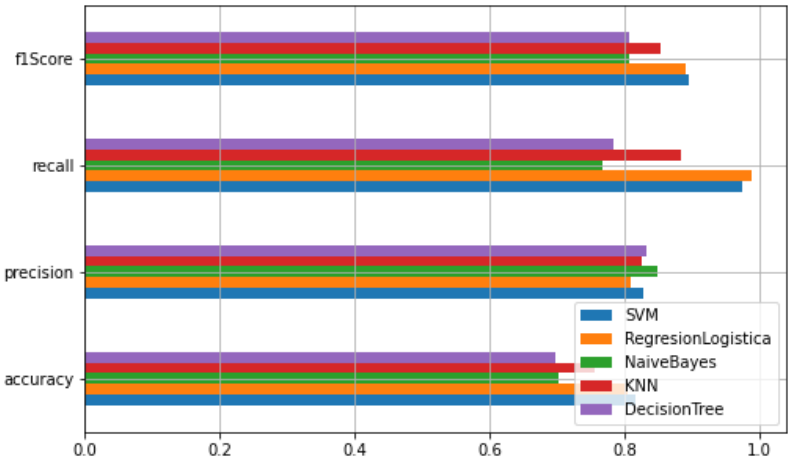


Figura 3.3. Valores de métricas para clasificador por valor.

En la Figura 3.3 se observa que los resultados de recall y f1 score son aceptables para regresión logística y SVM. Sin embargo, podríamos mejorarlos para verificar que modelo de clasificación es el adecuado y con que parámetros deberíamos entrenarlo.

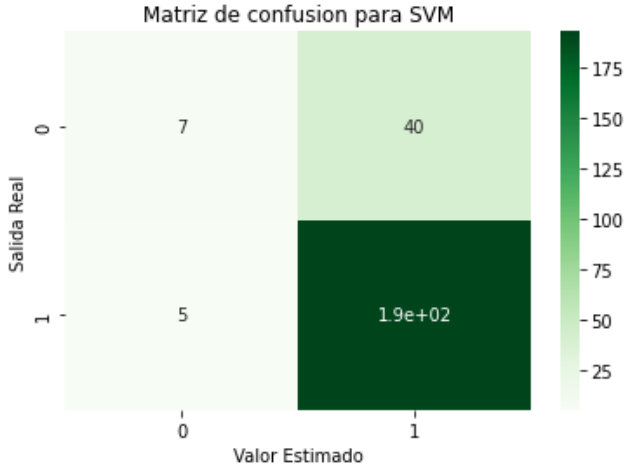


Figura 3.4. Matriz de confusión obtenida con modelo SVM para clasificador por valor.

La matriz de confusión que se presentan en la Figura 3.4 muestra que los datos a clasificar tienen mayor densidad que los que no tienen que ver con una compartición de información personal.

3.1.2 Resultados obtenidos con dataset balanceado

El problema del balanceo de los datos lo podemos combatir con el uso del parámetro “class_weight” dentro de los parámetros de cada clasificador. A continuación, se presenta los resultados obtenidos luego de este cambio.

- **Primer enfoque**

Los resultados obtenidos son mejor que los obtenidos con un dataset desbalanceado, lo cual era de esperarse. La Figura 3.5 muestra la gráfica de los valores obtenidos para las distintas métricas y los diversos clasificadores.

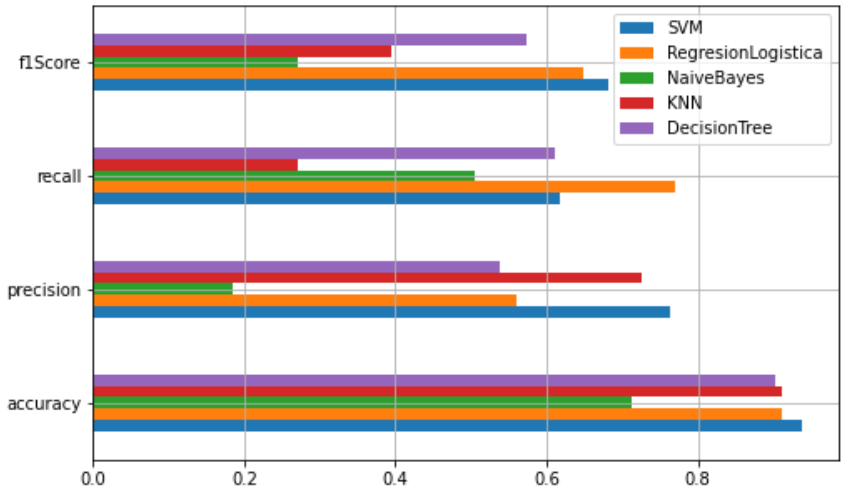


Figura 3.5. Métricas para clasificador general con dataset balanceado.

Si nos fijamos en los datos de recall vemos que incrementa en gran magnitud tanto para SVM, árbol de decisiones y regresión logística, siendo este último el mayor de todos. Con la Tabla 3.3 podemos corroborar los valores obtenidos para este caso de estudio; en el cual se observa que un valor aproximado del 77% para el recall es el mejor que se haya obtenido hasta el momento.

Tabla 3.3. Valores de métricas para clasificador general con dataset balanceado.

Métricas	SVM	R. Logística	Naive Bayes	KNN	Decision Tree
Exactitud	93.90%	91.13%	71.27%	91.18%	90.34%
Precisión	76.30%	55,93%	18.55%	72.50%	53.90%
Recall	61.68%	77.10%	50.46%	27.10%	61.21%
F1 Score	68.21%	64.83%	27.13%	39.45%	57.33%

El concluir que regresión logística sería el mejor modelo para clasificar una transferencia de datos aún es apresurado, puesto que el valor de f1 score es más bajo que el obtenido con SVM. Esto nos dice que el modelo no es preciso todavía. Tratar de mejorar estos valores nos permitirá conocer qué modelo es el adecuado para la clasificación de segmentos que tengan una etiqueta de transferencia de datos.

- **Segundo enfoque**

La Figura 3.6 muestra que los resultados en si no varían mucho al balancear el dataset.

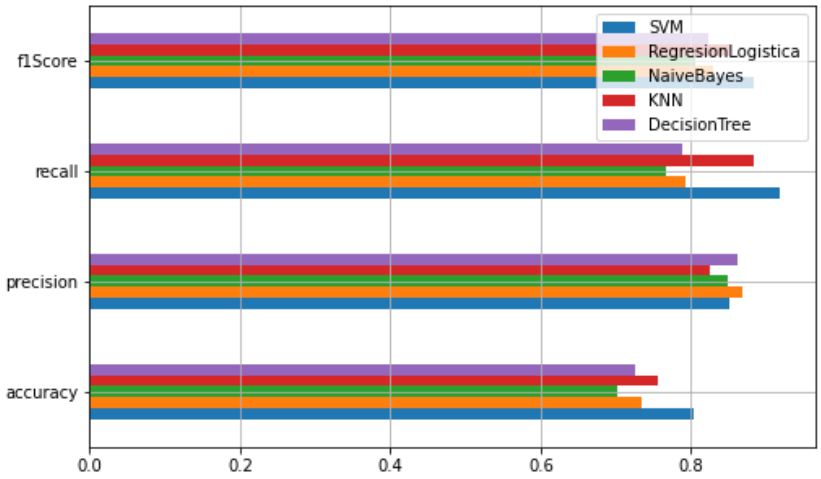


Figura 3.6. Métricas para clasificador por valor con dataset balanceado.

Vemos que los valores de recall y f1 score siguen siendo altos para todos los clasificadores, lo cual dice que estos detectan correctamente cuando se comparte información personal en los distintos segmentos ingresados. La Tabla 3.4 muestra los valores obtenidos en cada métrica, así se puede tener una mejor idea de que es lo que sucede.

Tabla 3.4. Valores de métricas para clasificador por valor con dataset balanceado.

Métricas	SVM	R. Logística	Naive Bayes	KNN	Decision Tree
Exactitud	80.40%	73.46%	70.20%	75.51%	72.65%
Precisión	85.04%	86.74%	84.91%	82.54%	86.18%
Recall	91.91%	79.29%	76.76%	88.38%	78.78%
F1 Score	88.34%	82.84%	80.63%	85.36%	82.32%

Al momento de balancear el conjunto de datos baja el recall para SVM. El cambo de ciertos parámetros puede hacer que estos valores aumenten o disminuyan por lo que es importante verificarlo. Hasta este punto los modelos de SVM y KNN son los de mejor rendimiento.,

3.1.3 Selección de mejores clasificadores

Para tratar de mejorar los resultados antes obtenidos se modifica el número de n-gramas, además de aumentar el parámetro de k-folds que permitirá tener un mejor procesamiento de los datos. El objetivo principal del uso de n-gramas es el de dar una cierta estructura al texto ingresado. Con esto se tendría un mayor entendimiento de las palabras y se podría clasificar de mejor manera los segmentos proporcionados.

- **Primer enfoque**

El conjunto de resultados obtenidos con ayuda del script descrito en la sección 2.2.8 da una perspectiva de cuál sería el mejor clasificador de segmentos en los que exista una anotación de transferencia de datos (ver enlace 4 Anexo VI). Estos resultados son alentadores dado que se obtiene métricas mayores al 70%. En este punto la observación del valor de recall es de gran importancia puesto que verifica con que modelo se tuvo un mayor porcentaje de datos identificados correctamente en la clase perteneciente. Dicho esto, se denota que el mejor valor de métricas se tiene con el clasificador de regresión logística y ciertos parámetros que se presentan en la Tabla 3.5

Tabla 3.5. Valores de métricas y parámetros del mejor modelo para clasificador general.

Precisión	Recall	F1-score	Accuracy	Parámetros
70.28%	74.47%	72.31%	85.42%	{'k-fold': 5, 'ngram-min': 1, 'ngram-max': 6, 'tf-idf': True, 'binary': True, 'classifier': 'RegresionLogistica'}

El recall aproximado del 74% y la precisión del 70% son los mejores valores que se obtuvieron durante el análisis elaborado para el presente trabajo. Por lo tanto, el mejor modelo para clasificar segmentos con una anotación de transferencia de datos.

- **Segundo enfoque**

La Tabla 3.6 muestra un resumen de los mejores modelos identificados por cada valor de elemento de transparencia.

Tabla 3.6. Valores de métricas y parámetros para mejores modelos por valor.

Valor	Precisión	Recall	F1-score	Accuracy	Parámetros
Sí comparte	95.00%	100.00%	97.43%	60.83%	{'k-fold': 10, 'ngram-min': 1, 'ngram-max': 4, 'tf-idf': False, 'binary': True, 'classifier': 'RegresionLogistica'}
Comparte información personal	84.00%	97.70%	90.30%	56.14%	{'k-fold': 20, 'ngram-min': 1, 'ngram-max': 6, 'tf-idf': False, 'binary': True, 'classifier': 'RegresionLogistica'}

Comparte a un tercero designado	80.75%	94.35%	86.91%	63.22%	{'k-fold': 10, 'ngram-min': 1, 'ngram-max': 2, 'tf-idf': True, 'binary': True, 'classifier': 'RegresionLogistica'}
Información recibe de la misma App	85.27%	98.58%	91.43%	58.29%	{'k-fold': 3, 'ngram-min': 1, 'ngram-max': 4, 'tf-idf': False, 'binary': True, 'classifier': 'RegresionLogistica'}
Información identifica al usuario	74.83%	99.03%	85.24%	56.27%	{'k-fold': 5, 'ngram-min': 1, 'ngram-max': 2, 'tf-idf': False, 'binary': True, 'classifier': 'SVM'}

Los valores obtenidos son satisfactorios y los parámetros utilizados se los observa en la última columna. Con respecto al uso de la función TF IDF en algunos casos puede ser útil dado que proporciona la relevancia de una palabra en un texto.

En la mayoría de los casos el mejor clasificador es regresión logística y solamente en uno de ellos es factible usar SVM que junto con los parámetros específicos se obtiene los mejores modelos para clasificar cada valor escogido. Los resultados obtenidos luego de la ejecución del script de la sección 2.2.8 se lo puede observar en los enlaces 4-8 del Anexo IV.

Con los algoritmos antes seleccionados se construye el modelo ensamblado cuyos resultados se muestran en la siguiente sección.

3.1.4 Resultados obtenidos con ensamble de modelos clasificatorios

Luego de procesar el conjunto de validación descrito en la sección 2.2.9 con el modelo ensamblado se tiene distintos resultados que se presentan a continuación.

Clasificador transferencia de datos/Sí comparte información

La matriz de la Figura 3.7 muestra la distribución de las predicciones obtenidas para este clasificador.

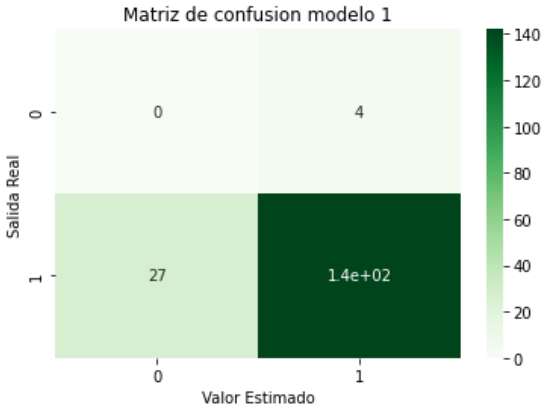


Figura 3.7. Matriz de confusión para modelo ensamblado 1.

Podemos ver que existe mayor distribución de segmentos en los cuales se tiene una anotación de transferencia y el valor de si comparte información lo cual se denota en la parte con el de mayor intensidad. No se identifica datos negativos, pero se tiene algunos que se identifican de manera errónea que en total serían 31 segmentos mal clasificados. Con estos datos se puede calcular las métricas de precisión, recall y f1 score, la Tabla 3.7 muestra los resultados obtenidos.

Tabla 3.7. Valores de métricas para modelo ensamblado 1.

Métrica	Valor
Precisión	97,26%
Recall	84,02%
F1	90,16%

.**Clasificador transferencia de datos/Recibe directamente de la aplicación**

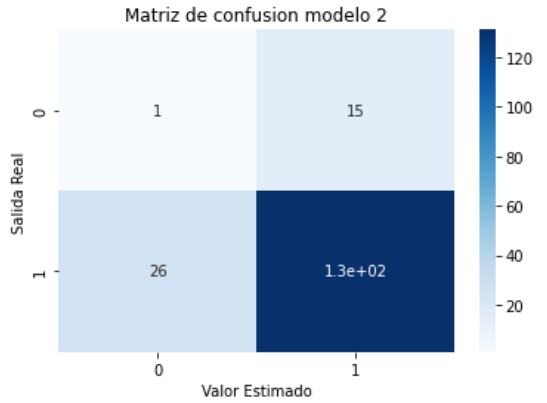


Figura 3.8. Matriz de confusión para modelo ensamblado 2.

La matriz de la Figura 3.8 muestra la clasificación obtenida con el segundo modelo de clasificación. Aquí se nota que se clasifica correctamente 133 segmentos en tanto que 41 se

los hace de manera errónea. Los resultados de precisión, recall y f1 score se presenta en la Tabla 3.8

Tabla 3.8. Valores de métricas para modelo ensamblado 2.

Métrica	Valor
Precisión	89,73%
Recall	83,44%
F1	86,47%

Clasificador transferencia de datos/Recibe directamente de la aplicación

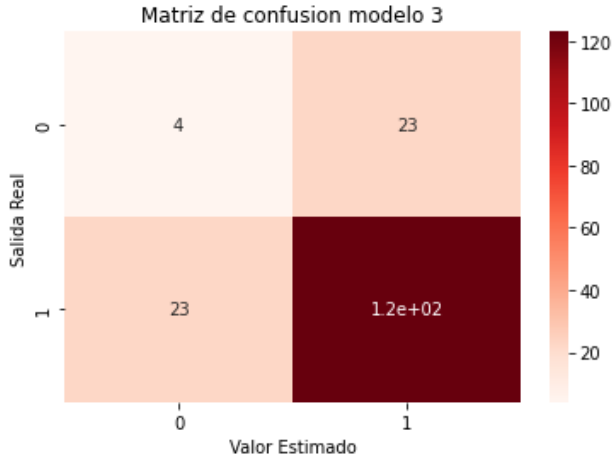


Figura 3.9. Matriz de confusión para modelo ensamblado 3.

En este caso se clasifican correctamente 127 segmentos en tanto que 46 segmentos no se los identifica de manera correcta. Los resultados de precisión, recall y f1 score se muestran en la Tabla 3.9

Tabla 3.9. Valores de métricas obtenidas para modelo ensamblado 3.

Métrica	Valor
Precisión	84,25%
Recall	84,25%
F1	84,25%

Los resultados denotan que el clasificador trabaja de manera eficiente dado que identifica los segmentos que tienen relación con una anotación de transferencia y con el valor especificado. El 84.25% en recall, precisión y f1 score indica un buen rendimiento de este modelo.

Clasificador transferencia de datos/Información personal

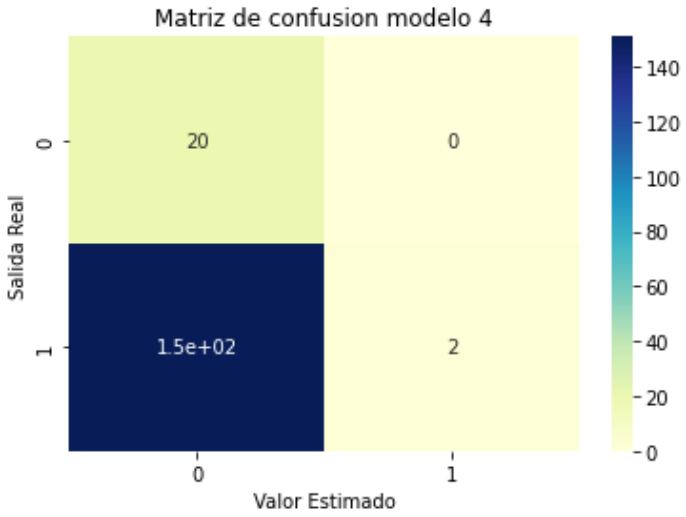


Figura 3.10. Matriz de confusión obtenida con modelo ensamblado 4.

La clasificación de este modelo en particular es deficiente, la matriz de confusión de la Figura 3.10 muestra que se clasifican correctamente 22 segmentos de los cuales 2 de ellos se tiene una anotación de transferencia de datos y el valor de comparte información personal. 151 segmentos se los clasifica de manera errónea aludiendo que no tiene dichas características, presumiblemente, esto se debe a la falta de datos dentro del entrenamiento de los modelos individuales o simplemente que el conjunto de prueba no cuenta con los datos necesarios para su clasificación. La Tabla 3.10 muestra los valores de recall, precisión y f1 score obtenidos en los cálculos.

Tabla 3.10. Valores de métricas para modelo ensamblado 4.

Métrica	Valor
Precisión	100,00%
Recall	1,31%
F1	2,58%

Clasificador transferencia de datos/Identifica al usuario

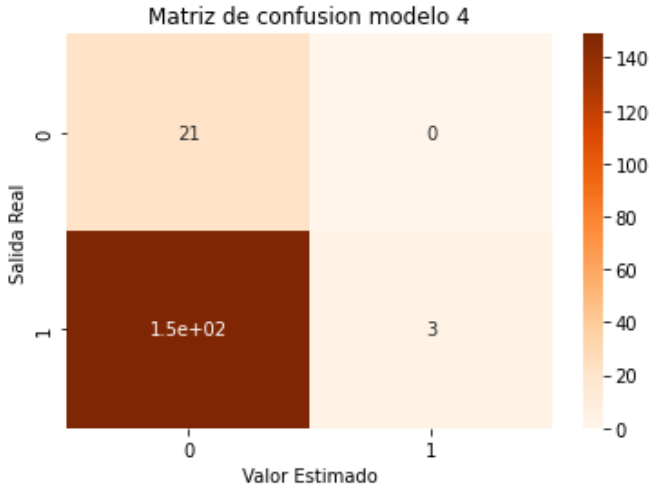


Figura 3.11. Matriz de confusión para modelo ensamblado 5.

De la misma manera que el modelo anterior la clasificación es deficiente. Vemos en la Figura 3.11 que se clasifica correctamente 24 segmentos de los cuales 3 tienen relación con transferencia y el valor de identifica al usuario. De la misma manera, puede deberse a la falta de datos en el entrenamiento de los modelos individuales o la falta de datos dentro del dataset de validación. La Tabla 3.11 muestra los valores de métricas obtenidos con este modelo.

Tabla 3.11. Valores de métricas obtenidas para modelo ensamblado 5.

Métrica	Valor
Precisión	100,00%
Recall	1,97%
F1	3,87%

3.2 Conclusiones

- En el presente trabajo se ha seguido un método sistemático para construir un etiquetador de transferencias de datos en políticas de privacidad en español. Primero, se ha construido un dataset con anotaciones de transferencias en políticas de privacidad descargadas de Play Store. Segundo, se ha entrenado un modelo de clasificación binario usando algoritmos de ML y PLN. Este se ha ensamblado con clasificadores de elementos de transparencia.

- Se ha conseguido desarrollar modelos con un performance mayor al 80% para identificar una práctica de transferencia de datos en políticas de privacidad en español. Tanto regresión logística como SVM demuestran ser los algoritmos adecuados para este propósito.

- En el caso del algoritmo Naive Bayes se tiene valores de 0% en las distintas métricas (ver Enlace 3 del Anexo IV). El problema puede deberse a la cantidad de datos negativos dentro del dataset, es decir, aquellos que no tienen una anotación de transferencia de datos. En cierta manera esto ocasiona que el algoritmo GaussianNB falle e identifique solo los datos que no tengan una anotación de transferencia de datos. El algoritmo KNN también tiene valores bajos de métricas.

3.3 Recomendaciones

- Para el mejoramiento de los distintos modelos elaborados se podría utilizar otras técnicas de preprocesamiento de datos.
- Contar con una gran diversidad políticas de privacidad haría que los modelos clasificatorios mejoren su rendimiento. En este trabajo se ha empleado un conjunto de 100 políticas, que podría ser extendido en trabajos futuros.
- Se sugiere construir clasificadores para los elementos de transparencia no implementados en este trabajo. Se podría anotar un conjunto más amplio de políticas de privacidad.

4 REFERENCIAS BIBLIOGRÁFICAS

[1] Asamblea Nacional del Ecuador, “Ley orgánica de protección de datos personales”, Quito, 2021.

[2] L. Enríquez, “Paradigmas de la protección de datos personales en Ecuador. Análisis del proyecto de Ley Orgánica de Protección a los Derechos a la Intimidad y Privacidad sobre los Datos Personales”, UASB, Quito, 2017.

[3] F. Miño, “Desarrollo de una herramienta web para la anotación de tratamientos de datos personales en políticas de privacidad en español”, Escuela Politécnica Nacional, Quito, 2021.

[4] H. Nissenbaum, “Privacy as Contextual Integrity”, Symposium, Washington, 2004.

[5] N. S. Amaya, “La privacidad como integridad contextual y su aplicación a las redes sociales”, Zer - Revista de Estudios de Comunicación, vol. 20, nº 39, 2015.

[6] M. Hernández y J. Gómez, “Aplicaciones de Procesamiento de Lenguaje Natural” Revista Politécnica, Quito, 2013.

[7] S. García, S. Ramírez, J. Luengo y. F. Herrera, “Big Data: Preprocesamiento y calidad de datos” julio 2016. [En línea]. Available:

https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf.

[8] D. Guamán, “Automated annotation of natural language privacy policies”, Madrid, 2020.

[9] A. Ribas, “Aprendizaje automático”, Edicions UPC, Barcelona, 1994.

[10] L. Sandoval, “Algoritmos de aprendizaje automático para análisis y predicción de datos.”, Revista tecnológica, nº 11, pp. 37-38, 2018.

[11] E. Menasalvas, A. Rodríguez, M. Guzmán, S. Jiménez y S. Duque, “Algoritmos de machine learning”, ManagementSolutions, Madrid, 2021.

[12] J. Heras, “IArtificial.net”, 09 octubre 2020. [En línea]. Available: https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/.

[13] “statdeveloper”, [En línea]. Available: https://www.statdeveloper.com/evaluacion-del-modelo-de-clasificacion/#:~:text=Matr%C3%ADz%20de%20confusi%C3%B3n%20(F1%2Dscore, etiquetas%20predichas%20por%20el%20clasificador.

[14] E. Loper, “NLTK: The Natural Language Toolkit”, CoRR, cs.CL/0205028. 10.3115/1118108.1118117, 2022.

[15] Scikit-learn, “scikit-learn” [En línea]. Available: https://scikit-learn.org/stable/getting_started.html.

[16] Google, “Colaboratory”, [En línea]. Available: https://research.google.com/colaboratory/intl/es/faq.html#:~:text=Colaboratory%2C%20o%20%22Colab%22%20para,an%C3%A1lisis%20de%20datos%20y%20educaci%C3%B3n..

[17] National Science Foundation, “usableprivacy”, 2015. [En línea]. Available: https://usableprivacy.org/.

[18] Javatpoint, “Support Vector Machine Algorithm”, [En línea]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[19] Javatpoint, “Logistic Regression in Machine Learning”, [En línea]. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning.

[20] Javatpoint, “Decision Tree Classification Algorithm”, [En línea]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[21] Javatpoint, “Naïve Bayes Classifier Algorithm”, [En línea]. Available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

[22] Javatpoint, “K-Nearest Neighbor(KNN) Algorithm for Machine Learning”, [En línea]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[23] C. Martínez, “Máquinas de vectores de soporte”, junio 2018. [En línea]. Available: https://rpubs.com/Cristina_Gil/SVM.

[24] J. López, “Análisis y uso de funciones de base radial como filtros interpoladores”, febrero 2009. [En línea]. Available: https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/360/1/AguilarLJC.pdf.

[25] M. Quintana, J. Flores, S.Salas y V. Landassuri, “Ensamble de clasificadores para determinar el perfil académico del estudiante usando árboles de decisión y redes neuronales”, 2018. [En línea]. Available: https://rcs.cic.ipn.mx/2018_147_5/Ensamble%20de%20clasificadores%20para%20determinar%20el%20perfil%20academico%20del%20estudiante%20usando%20arboles.pdf.

5 ANEXOS

ANEXO I. Aplicaciones utilizadas para estudio

ANEXO II. Diagrama de flujo de procesamiento de etiquetado

ANEXO III. Diagrama de flujo de procesamiento de etiquetado actualizado

ANEXO IV. Enlaces

ANEXO I. Aplicaciones utilizadas para estudio

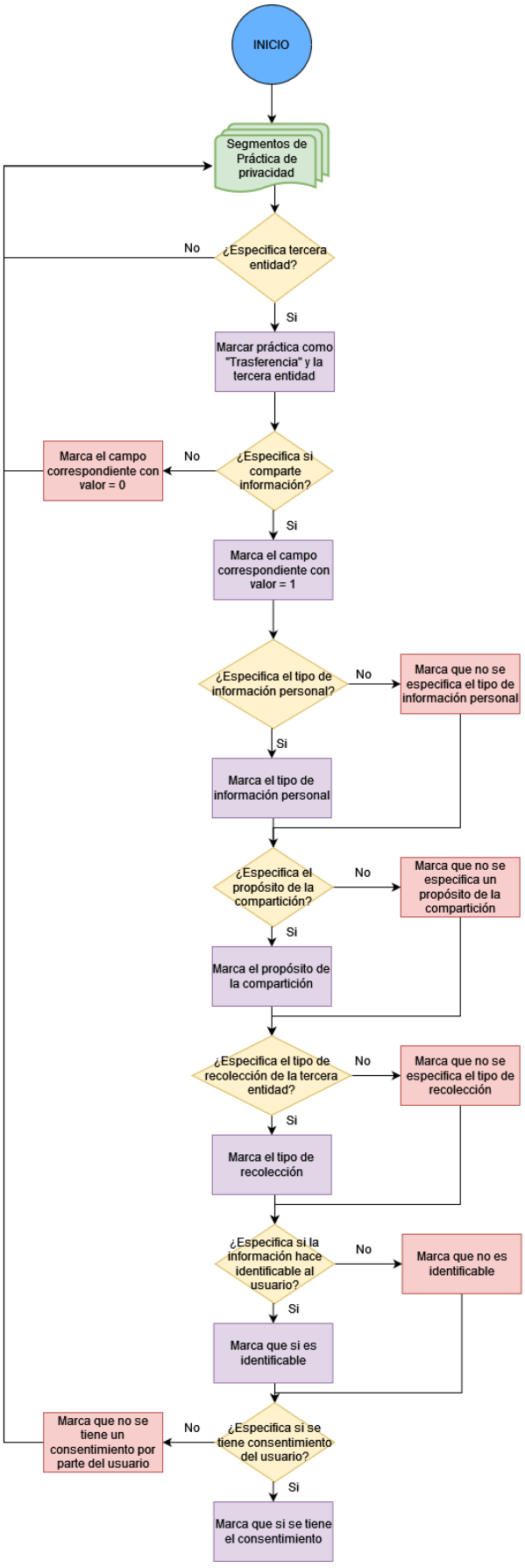
Politica	**Categoría de la política**	**País**
Adidas	Deportes	Alemania
Adobe	Herramientas	EEUU
Airbnb	Viajes	EEUU
AliExpress	Compras	China
Amazon	Compras	EEUU
ANT	Herramientas	Ecuador
App canton Mejía	Herramientas	Ecuador
App Latacunga movil	Herramientas	Ecuador
B612	Fotografía	Corea del Sur
Banco Guayaquil	Banca	Ecuador
Banco Internacional	Banca	Ecuador
Banco Pacifico	Banca	Ecuador
Banco Pichincha	Banca	Ecuador
Banco Produbanco	Banca	Ecuador
Booking	Viajes	EEUU
Busuu	Educación	España
Cabify	Mapas y Navegación	España
CamScanner	Herramientas	China
Cap Cut	Herramientas	China
CineXt	Entretenimiento	Ecuador
Coursera	Educación	EEUU
Crunchyroll	Entretenimiento	EEUU
Cuenca en linea	Herramientas	Ecuador

DeUna!	Banca	Ecuador
DGO app	Entretenimiento	EEUU
Diners Ecuador	Banca	Ecuador
Discord	Social	EEUU
Disney	Entretenimiento	EEUU
Dominos Ecuador	Comida y Bebidas	Ecuador
Dropbox	Herramientas	EEUU
Duolingo	Educación	EEUU
ecu911	Herramientas	Ecuador
ElCanalDelFutbol	Entretenimiento	Ecuador
Emaseo	Herramientas	Ecuador
ESPN	Entretenimiento	EEUU
Evernote	Herramientas	EEUU
Facebook	Social	EEUU
Formula 1	Herramientas	EEUU
Github	Productividad	EEUU
Glovo	Comida y Bebidas	España
Gob.ec	Herramientas	Ecuador
Google	Herramientas	EEUU
HBOMax	Entretenimiento	EEUU
ILovePDF	Herramientas	España
InDrive	Mapas y Navegación	EEUU
InShot	Fotografía	EEUU
Instagram	Social	EEUU
KFCappEC	Comida y Bebidas	Ecuador
Linkedin	Social	EEUU
Luminosity	Productividad	EEUU
MalwareBytes	Herramientas	EEUU
McDonalds	Comida y Bebidas	EEUU
MercadoLibre	Compras	Argentina
Mi Volkswagen	Autos	Ecuador
Microsoft	Herramientas	EEUU
Microsoft Edge	Herramientas	EEUU
Multicines	Entretenimiento	Ecuador
MyKia	Autos	Ecuador

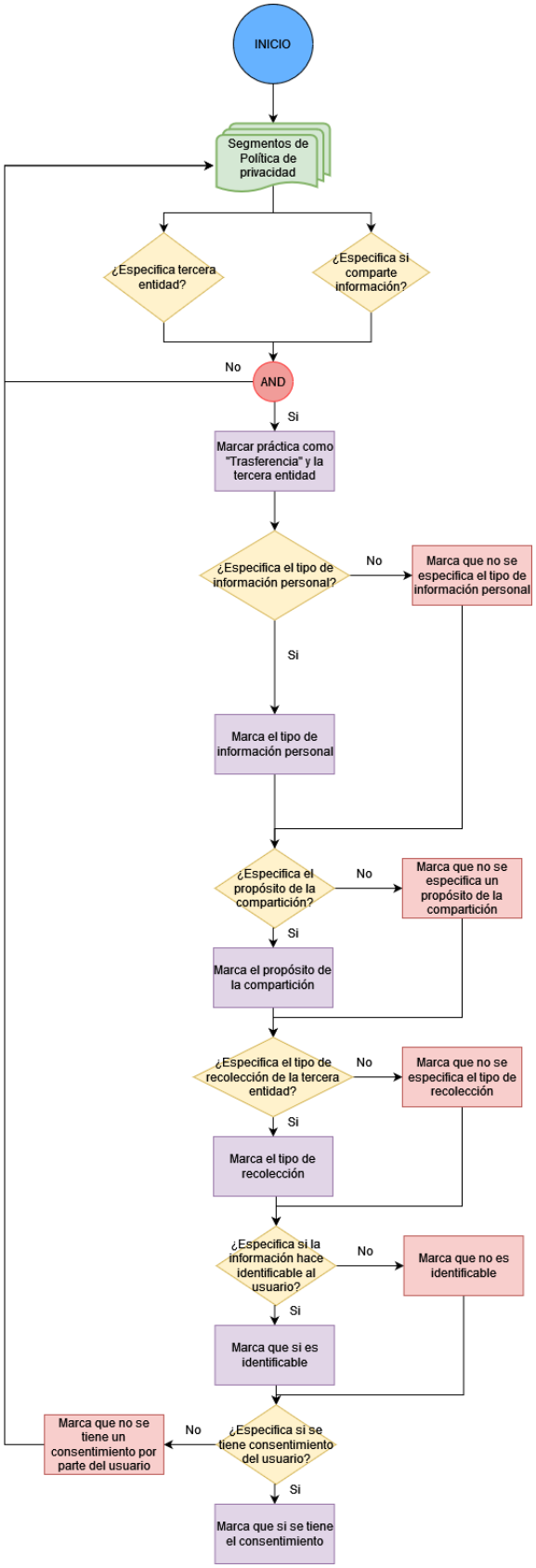
netflix	Entretenimiento	EEUU
Nextdoor	Social	EEUU
Nike	Deportes	EEUU
Opera	Herramientas	Noruega
Paramount+	Entretenimiento	EEUU
PatioTuerca	Compras	Ecuador
PayPal	Banca	EEUU
Pedidosya Ecuador	Compras	Ecuador
PhotoRoom	Fotografía	Francia
Pinterest	Social	EEUU
PizzeriaHornero	Compras	Ecuador
Platzi	Educación	Colombia
Prime Video	Entretenimiento	EEUU
Reddit	Social	EEUU
Retrica	Fotografía	Corea del Sur
Shappi	Compras	EEUU
Shein	Compras	España
Slack	Social	EEUU
Snapchat	Social	EEUU
SoundHound	Música y Audio	EEUU
SRI	Herramientas	Ecuador
StarPlus	Entretenimiento	EEUU
StartMaker	Social	EEUU
TaDa	Comida y Bebidas	Ecuador
Telegram	Social	Dubai
TiendaMia	Compras	EEUU
TikTok	Social	China
Tinder	Citas	EEUU
Tipti	Compras	Ecuador
Toomics	Entretenimiento	Corea del Sur
Trivago	Viajes	Alemania
Tubi	Entretenimiento	EEUU
Twitch	Social	EEUU
Twitter	Social	EEUU
Uber	Mapas y Navegación	EEUU

Udemy	Educación	EEUU
Vitality	Productividad	Ecuador
Waze	Mapas y Navegación	Israel
Whatsapp	Social	EEUU
Whatsapp Business	Social	EEUU
Yahoo	Herramientas	EEUU
Zoom	Social	EEUU
TuneIn	Música y Audio	EEUU
Plex	Entretenimiento	EEUU
Deezer	Música y Audio	Francia
Babbel	Educación	Alemania
Brainly	Social	EEUU

ANEXO II. Diagrama de flujo de procesamiento de etiquetado



ANEXO III. Diagrama de flujo de procesamiento de etiquetado actualizado



ANEXO IV. Enlaces

Enlace 1. Enlace a conjunto de anotaciones utilizado.

https://docs.google.com/spreadsheets/d/1R2RFZUTQ7ex6pST785MDkMWjwBS0261RLgAD3RKsfp0/edit?usp=sharing

Enlace 2. Enlace a carpeta de archivos YAML.

https://drive.google.com/drive/u/0/folders/1aB9g3IlzRrn4pqbrB9jMEBqirt7zIVCY

Enlace 3. Enlace a archivo de resultados de mejor modelo para clasificador general.

https://docs.google.com/spreadsheets/d/1J2Ki1k1N9JJKY84VWq5LuMyYEOS__DF8mw4LrayUh3M/edit?usp=sharing

Enlace 4. Enlace a archivo de resultados para valor sí comparte

https://docs.google.com/spreadsheets/d/1sTZ_DR7oQd5qz-SAMbTO2ugHHCQgXaeXY0ADycSaW8I/edit?usp=sharing

Enlace 5. Enlace a archivo de resultados para valor recibe directamente de la aplicación

https://docs.google.com/spreadsheets/d/1X6vSS6plm9IImzBFelheukj0ZsoyB-ZqYf1LK-VbIoQ/edit?usp=sharing

Enlace 6. Enlace a archivo de resultados para valor tercero designado

https://docs.google.com/spreadsheets/d/1pGPTrscHl0Ve7I8IalK76Gn25ntE80d-EFlED1X1BxA/edit?usp=sharing

Enlace 7. Enlace a archivo de resultados para valor información personal

https://docs.google.com/spreadsheets/d/1ArIf2CIriskhvFUphA-_LvbwpoJaklIhXXfXwoPFIFM/edit?usp=sharing

Enlace 8. Enlace a archivo de resultados para valor identifica al usuario

https://docs.google.com/spreadsheets/d/1MX_MnWv5jhFIW1DSRsStreLGEkuYXxPVcJij4Q5CMGE/edit?usp=sharing