CAPÍTULO 3. FLUJO DE TRABAJO PARA LA INTEGRACIÓN EN LOS ESTÁNDARES FEDERADOS

En este capítulo se propone un flujo de trabajo, que cualquier RP puede implementar, para integrar las técnicas de análisis de comportamientos en los esquemas de gestión de identidades federados (p. ej. OAuth y OIDC) [110]. Tal y como se ha podido ver en el capítulo del estado del arte (Sección 2.1), en las especificaciones federadas, las credenciales de los usuarios son almacenadas por el IdP. Cuando un usuario trata de acceder a un recurso, servicio o aplicación (es decir, SP o en el entorno federado RP), la RP confía en el IdP para solventar los procesos de IAAA. De este modo, el usuario final se autentica de forma externa a la RP, es decir, en el IdP, obteniendo una acreditación en forma de *token*. Finalmente, el usuario final utiliza este *token* para comunicarse con la RP, y poder así interactuar con ella.

La solución propuesta en esta tesis doctoral para mejorar la seguridad de estos esquemas federados se basa en utilizar los modelos de análisis de comportamientos. Cuando una RP utiliza un flujo federado para solventar los procesos de IAAA, está delegando parte de la seguridad en el IdP. Sin embargo, delegar los procesos de IAAA en un agente externo no tiene por qué significar delegar todos los aspectos de la seguridad como, por ejemplo, desde el punto de vista de la prevención o de la detección.

La RP tiene un rol muy significativo a la hora de proteger a sus usuarios de recibir un ciberataque como, por ejemplo, ataques de suplantación de identidad, ya que los usuarios interactúan la mayor parte del tiempo con ella. Cuando se utiliza un esquema de gestión de identidades federado, la mayoría de las partes implicadas asumen que la seguridad pasa a ser responsabilidad del IdP en exclusiva, siendo esta una suposición errónea. Tanto IdP como RP deben estar concienciados en asumir su responsabilidad en cuanto a la seguridad, para poder proteger a los usuarios de las posibles amenazas que existen.

De aquí en adelante, se detalla el flujo de trabajo propuesto, que puede utilizar cualquier RP con el objetivo de mejorar los niveles de seguridad proporcionados al usuario final. Este flujo de trabajo se centra en la prevención y detección de ataques de suplantación de identidades por medio del robo de credenciales o secuestros de sesión. Se basa en utilizar técnicas de análisis de comportamiento, y en los detalles de integración con los principales estándares de gestión de identidades.

3.1. Arquitectura y premisas

A lo largo de este capítulo se asume una arquitectura federada en la que existen tres roles principales: el EU, el IdP y la RP.

Las RPs necesitan conocimiento acerca del comportamiento de sus usuarios. Este conocimiento se suele extraer de diferentes tecnologías, interfaces o APIs que son capaces de recopilar información relativa al software o hardware utilizado por el EU, aspectos de configuración (ya sea del sistema o del cliente web utilizado) y de su comportamiento. Toda esta información se utiliza para generar una huella digital, capaz de identificar única y exclusivamente a cada EU.

Una huella digital puede suponer una invasión para la privacidad de los usuarios de los que se recopila información. Estas huellas se pueden utilizar para fines como el perfilado de usuario o para la personalización de contenido publicitario, lo cual puede suponer un rechazo por parte de los usuarios y, por consiguiente, la negativa a su adopción [111]. Hoy en día, existen multitud de trabajos de investigación y de productos finales de empresas privadas que protegen a los clientes web y a las aplicaciones de los usuarios, bloqueando los códigos encargados de la recopilación de la información o cambiando los valores de los atributos que estos recopilan [112].

Por todo lo expuesto anteriormente, en este trabajo se asumen las siguientes premisas:

- Existe suficiente diversidad en el comportamiento de los usuarios como para poder generar una huella digital, única y identificativa para cada usuario.

- La RP no colabora con el IdP para generar esta huella, ni para analizarla y detectar posibles anomalías que suponen una brecha de seguridad para la RP. Esto quiere decir que, la RP es capaz de seguir el flujo de trabajo propuesto de forma independiente y utilizando única y exclusivamente sus propios recursos.

- La RP informa a los usuarios del flujo de trabajo que va a utilizar con el objetivo de mejorar los niveles de seguridad ofrecidos. Más concretamente, la RP informa a los usuarios que este flujo de trabajo puede prevenir y detectar ataques de robo de credenciales y de secuestro de sesión.

- La RP cumple con las garantías de privacidad y protección de datos (Reglamento General de Protección de Datos (RGPD) y Ley Orgánica de Protección de Datos de Carácter Personal (LOPD)). Además, sigue los principios de privacidad desde el diseño minimizando la recogida de información,

obteniendo el necesario consentimiento explícito e informado del usuario final, y garantizando los niveles adecuados de transparencia.

3.2. Casos de uso

En general, el flujo de trabajo propuesto a continuación puede ser implementado por cualquier RP para mejorar los niveles de seguridad cuando se utiliza un estándar de gestión de identidades federado para realizar los procesos de IAAA. Más específicamente, este flujo de trabajo permite mejorar los niveles de seguridad añadiendo nuevos mecanismos para el control de accesos utilizando los sistemas basados en riesgos, la autenticación continua y los sistemas de alerta temprana. Se han seleccionado estos tres casos de uso representativos, en los que la implantación del flujo de trabajo es de utilidad:

- Caso de uso 1: una RP decide utilizar el nivel de garantía (en inglés, *Level of Assurance* [LoA]) en la petición de autenticación. De esta manera, utilizando el LoA, una RP puede especificar el grado de confianza que requiere al IdP para poder autenticar a un usuario (uno o dos factores de autenticación, factores biométricos o criptografía). Normalmente, el LoA se define de forma estática, sin embargo, gracias al uso del flujo de trabajo propuesto, este valor se puede fijar de forma dinámica en función de los valores devueltos por los modelos de análisis de comportamientos. Por ejemplo, un comercio electrónico en el que un usuario realiza una compra desde un navegador totalmente distinto al que suele utilizar, y sus dinámicas de comportamiento del uso del teclado y ratón son anómalas. En este caso, la RP puede solicitar al IdP un LoA mayor, de tal manera que el EU debe mostrar más autenticadores para terminar satisfactoriamente el proceso de autenticación. Este comportamiento anómalo es un signo de una posible suplantación de identidad. En caso de que el usuario sea legítimo podrá demostrarlo por medio de más autenticadores, mientras que para un atacante será más difícil o imposible.

- Caso de uso 2: una RP decide implantar un mecanismo de autenticación continua durante las sesiones de sus usuarios. En el flujo de trabajo propuesto, estos mecanismos de autenticación continua son implementados mediante modelos de análisis de comportamientos. De esta manera, el EU deberá de presentar los autenticadores necesarios, cuando la RP lo considere a lo largo de una sesión, teniendo en cuenta los comportamientos anómalos detectados por los modelos de análisis de comportamientos. Este proceso le permitirá quedar autenticado recibiendo los *tokens* pertinentes. Nuevamente, estos comportamientos anómalos pueden estar vinculados a un evento que no sea necesariamente un incidente de ciberseguridad, o por el contrario pueden ser debidos a un secuestro de sesión, el cual quedaría bloqueado gracias a esta propuesta.

- Caso de uso 3: una RP decide almacenar el histórico de comportamiento de todos los EU. En este sentido, el análisis de toda esta información se prepara para ser procesado de forma periódica (p. ej. cada noche o cada semana). Los modelos de análisis de comportamientos determinan si se han detectado comportamientos anómalos, de tal manera que se levanten alertas en caso afirmativo. De esta forma, tanto los EU como las RPs pueden tomar las acciones necesarias teniendo en cuenta estas alertas. Este procesamiento no se realiza en tiempo real, como sucede en los casos de uso 1 y 2. La gran ventaja de este caso de uso es que la no tener el requisito de procesar la información en tiempo real, los modelos de análisis de comportamientos tienden a ser más precisos y eficaces. Por otro lado, la desventaja es que la brecha de seguridad ya habrá sucedido y por tanto no se podrá evitar, únicamente se podrá levantar una alerta para tener constancia de su ocurrencia.

Tal y como se ha comentado en el estado del arte, la gestión de identidades no solo es aplicable a usuarios, sino también a entidades (p. ej. sensores en entornos IoT). Cabe destacar, que los casos de uso aquí propuestos también son de utilidad para esta casuística. Por ejemplo, la fase de alta o *enrollment* en sensores en una ciudad inteligente puede ser muy costosa debido a la gran escala de este tipo de proyectos. El análisis de comportamiento es de gran utilidad para solventar este caso de uso, así como sus fases posteriores de autenticación y autorización [113]. Otro caso de uso en este ámbito es la detección de intrusiones analizando el tráfico de red de las comunicaciones de los sensores [114].

3.3. Flujo de trabajo

El flujo de trabajo propuesto en esta tesis doctoral, proporciona a las RPs las líneas generales para integrar soluciones basadas en el análisis de comportamiento dentro de los esquemas de gestión de identidades federados. Este flujo está pensado para poder ser utilizado en todos los casos de uso detallados en la Sección 3.2. Estos casos de uso pueden ser previos o posteriores a la autenticación, algunos son *offline* y otros son *online*. Además, los recursos de computación o los datos recopilados, así como los consentimientos necesarios por parte de los EU son muy diversos.

En la Figura 3.1 se muestra el flujo general que han de seguir todos los agentes que interactúan en un flujo federado de gestión de identidades. Es decir, esta figura extrapola el funcionamiento de cualquier estándar de gestión de identidades federado, y el lugar donde se integran las técnicas de análisis de comportamiento y cada caso de uso. Estos pasos no están enumerados, pues se pueden realizar de forma asíncrona, es decir, los agentes pueden ir avanzando y volviendo a pasos anteriores en función del estado en el que se encuentren.

En la Figura 3.2 se detallan los pasos precisos que ha de seguir cualquier RP para integrar las técnicas de análisis de comportamientos dentro del flujo visto en la Figura 3.1. Consta de cinco pasos fundamentales: la selección de la huella digital, la generación de la huella digital, el modelado de las huellas digitales, la evaluación de los modelos generados y su integración en los flujos de información de los estándares federados. Cabe destacar, que una RP puede avanzar y retroceder en la realización de estos pasos en función de los resultados parciales obtenidos en los mismos con el objetivo de lograr una integración satisfactoria.

El primer paso es definir y seleccionar la huella digital. Para ello tiene que utilizar información de comportamiento y atributos lo suficientemente relevantes y descriptivos para que cada huella sea única. De esta forma, las huellas son distinguibles entre sí y por tanto se pueden detectar anomalías de comportamientos.

Figura 3.1. Visión global de la integración del análisis de comportamiento en un estándar federado

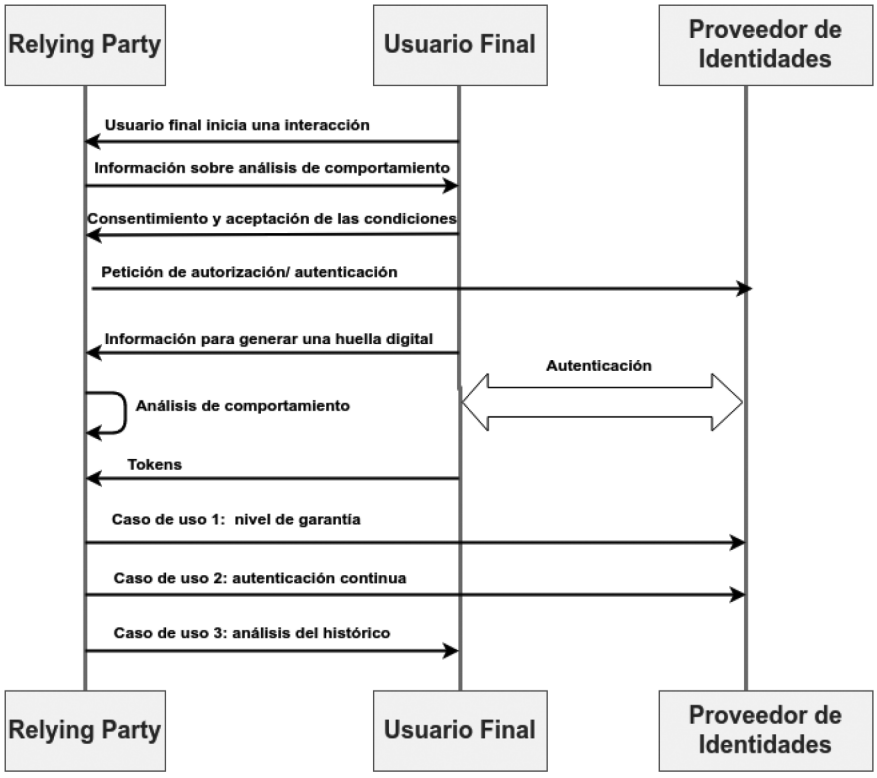
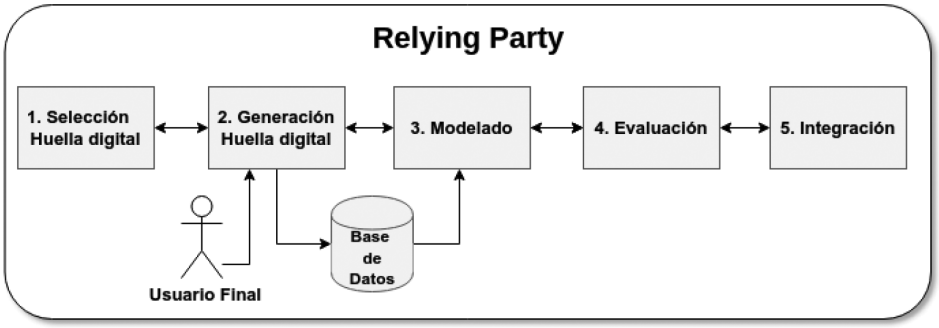


Figura 3.2. Flujo de trabajo para la integración del análisis de comportamiento en un estándar federado



El segundo paso se centra en generar esta huella digital considerando la información y atributos previamente seleccionados. Para ello se debe utilizar la información recopilada y limpiar la información no relevante (ruido), realizar un preprocesamineto de los datos y transformarlos en variables descriptivas que puedan ser utilizados por los algoritmos de aprendizaje máquina. Posteriormente, el conocimiento extraído en forma de variables descriptivas ha de almacenarse de forma eficiente en una base de datos robusta y escalable. Además, se han de definir procesos para verificar que las huellas digitales son válidas a lo largo del tiempo, ya que estas dinámicas y pueden ir cambiando y poder así actualizarlas.

El tercer paso es modelar las huellas digitales. Para ello se va a hacer uso de los modelos de análisis de comportamiento (es decir, modelos de aprendizaje máquina) específicos que permiten realizar la detección de anomalías. Estos algoritmos se nutren del conocimiento extraído en el segundo paso, por lo que, si este segundo paso no se realiza de forma óptima, es de esperar que el resultado obtenido para el tercer paso no sea tampoco óptimo. Los modelos de análisis de comportamientos tienen que permitir comparar nuevos comportamientos con las huellas digitales generadas, con el objetivo de poder encontrar anomalías y por consiguiente detectar posibles brechas de seguridad. Para obtener buenos resultados, una RP ha de considerar una métrica de similitud o distancias entre huellas digitales precisa. Posteriormente, ha de ser capaz de fijar un umbral de decisión a partir del cual una muestra analizada se considere legítima o una anomalía. Finalmente, al igual que con la generación de huellas digitales, los modelos de aprendizaje máquina han de estar preparados para reentrenarse con el objetivo de considerar los cambios de comportamiento intrínsecos que surgen en los propios usuarios. Esto último, también se puede realizar considerando un modelo de entrenamiento online, es decir, un modelo que va entrenándose a medida que va realizando nuevas predicciones en el sistema.

El cuarto paso es evaluar los modelos de análisis de comportamientos, elaborados en el paso anterior. Este proceso permite detectar posibles errores de diseño que no se han considerado en un primer momento como, por ejemplo, que las huellas digitales seleccionadas y generadas en los pasos previos, no son lo suficientemente descriptivas como para ser funcionales. Además, permite determinar si los modelos de aprendizaje máquina obtenidos se comportan de la manera esperada, es decir, si permiten detectar anomalías de comportamiento, son robustos y tienen capacidad de generalización. En caso de que no se comporten de la manera esperada, la RP ha de evaluar los motivos y solventarlos volviendo atrás a alguno de los pasos anteriores. Cabe destacar, que las métricas estándares de evaluación de modelos de aprendizaje máquina no son lo suficientemente descriptivas en el ámbito del análisis de comportamiento. Es por esto que se suelen utilizar métricas específicas para este dominio.

Finalmente, la quinta tarea es la integración del flujo de trabajo en los estándares de gestión de identidades federados. Este proceso depende específicamente del estándar donde se quiera integrar el flujo. En este trabajo se establecen la líneas y procedimientos que una RP ha de considerar como, por ejemplo, modificar lo menos posible los flujos de información ya determinados por los estándares y utilizar los propios mecanismos que estos estándares proveen para modificarlos en caso de ser necesario. Además, al realizar esta integración y teniendo en cuenta los aspectos relativos a la privacidad, la RP ha de informar a los usuarios de la recopilación y tratamiento de los datos que se van a realizar.

De aquí en adelante, se va a analizar cada una de estas tareas en detalle.

3.3.1. Selección de huella digital

En esta sección se proponen una serie de atributos que cualquier RP puede utilizar para generar una huella digital con el objetivo de poder completar el flujo de trabajo propuesto. Cabe destacar, que es imposible definir una huella digital general que sirva como solución a todas y cada una de las RPs. Esto se debe a que, existen multitud de RPs de naturaleza totalmente distinta (p. ej. aplicación web, aplicación móvil) y que poseen recursos totalmente diferentes y heterogéneos para contemplar multitud de casos de uso en distintos dominios. Es por esto que cada RP ha de decidir un mínimo de atributos que garantice los niveles de seguridad necesarios en función de sus necesidades.

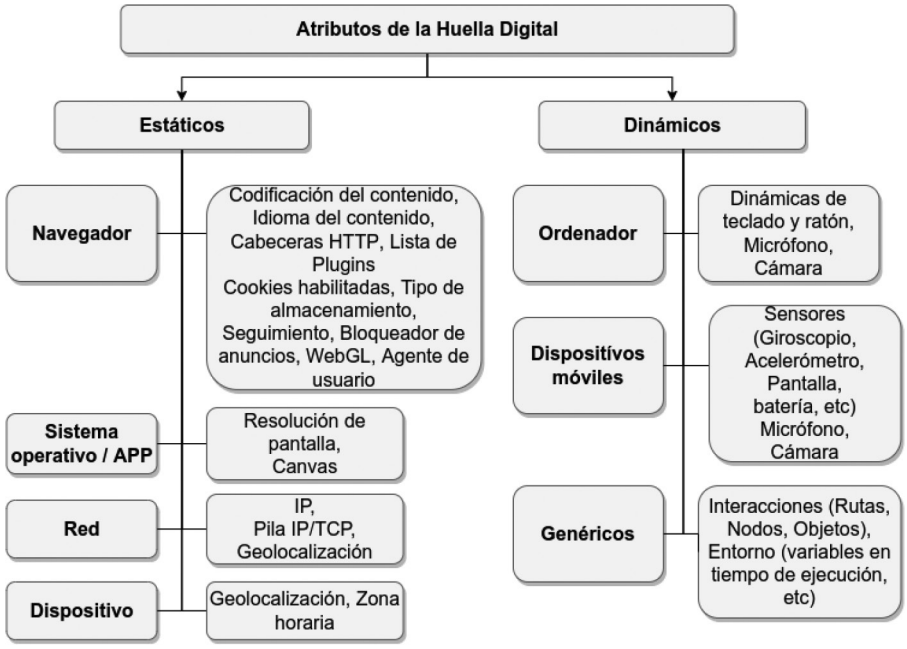
Debe existir una solución equilibrada entre eficacia y privacidad. Esto se debe a que, cuanto mayor sea el número de atributos seleccionados, más eficaces serán los modelos de análisis de comportamientos a la hora de detectar posibles brechas de seguridad. Sin embargo, a medida que el número de

atributos aumenta, el proceso de recolección de la información será más costoso y además será más invasivo para la privacidad del usuario final. Un claro ejemplo de este equilibrio se puede encontrar en la diferencia de aumentar la seguridad de una aplicación financiera y en una red social. En el caso de la aplicación financiera, posiblemente un usuario prefiera sacrificar privacidad a cambio de obtener un nivel de seguridad más alto, no siendo así para proteger su perfil en una red social. Toda RP dispuesta a integrar este flujo de trabajo ha de asegurarse el cumplimiento de la normativa sobre privacidad y protección de datos vigente.

Todos los atributos propuestos se pueden ver categorizados en la Figura 3.3. En primer lugar, estos atributos se dividen en estáticos y dinámicos. Los atributos estáticos están compuestos mayoritariamente por información de contexto. Estos atributos se consideran estáticos, porque suelen cambiar nada o muy poco a lo largo del tiempo. Por otro lado, los atributos dinámicos representan las interacciones del usuario final con la RP. Estos atributos son muy cambiantes a lo largo del tiempo. Muchos de los atributos estáticos están definidos por un conjunto finito de posibles valores, por lo que, pueden existir muchos usuarios con valores similares o idénticos, mientras que los atributos dinámicos pueden tomar valores muy heterogéneos (p. ej. no existe un patrón único de comportamiento en las dinámicas de teclado que agrupe un conjunto amplio de usuarios). El grupo de atributos estáticos que garantiza la suficiente singularidad en las huellas digitales, es decir, sus valores presentan la suficiente diversidad como para que la huella digital sea única, son los siguientes:

- Codificación del contenido: normalmente un algoritmo de compresión que el navegador soporta.
- Idioma del contenido: preferencias de lenguaje seleccionado por el usuario en el navegador.
- Cabeceras HTTP: parámetros contenidos en las cabeceras de las peticiones HTTP.
- Lista de plugins: plugins instalados en el navegador. Por ejemplo, *Java, Flash, o Silver- light*.
- *Cookies* habilitadas: configuración de *cookies* en el navegador.
- Tipo de almacenamiento: mecanismos de almacenamiento soportados por el navegador (local, *indexedDB*, sesión, *Web-SQL*, etc.).
- Seguimiento: cabecera que se incluye cuando no se desea obtener seguimiento de ciertas páginas o aplicaciones web.

Figura 3.3. Tipos de atributos de la huella digital



- Bloqueador de anuncios: software instalado que bloquea la publicidad agresiva o que afecta a la privacidad del usuario en el navegador.
- *WebGL*: *WebGL* es una API para renderizar los gráficos en el navegador. Esta API permite obtener diferentes atributos relacionados con el navegador instalado.
- Agente de usuario: información relacionada con el navegador y sistema operativo utilizado por el usuario.
- Resolución de pantalla, profundidad del color y densidad de píxeles: información relacionada con la tarjeta gráfica utilizada y los drivers instalados.
- Canvas: script que se utiliza para renderizar el texto y gráficos de HTML. Proporciona información del software y el hardware del usuario.
- Dirección IP: identifica la fuente de comunicación.
- TCP/IP: puertos abiertos, DNS utilizado, configuración NAT, etc.

- Geolocalización: obtenida por medio de inferencia por la red usada o por APIs específicas.

- Zona horaria: zona horaria seleccionada en la configuración del navegador.

La recomendación que se propone en esta tesis doctoral para seleccionar los atributos estátcos es la de seleccionar al menos un atributo de cada una de las categorías siempre y cuando sea posible.

Por otro lado, los atributos dinámicos están fuertemente ligados al dispositivo que se utiliza para acceder a la RP. Por ejemplo, las dinámicas de teclado, ratón o el uso de sensores de entrada como el micrófono o cámara, están muy ligados a dispositivos como un ordenador personal. Sin embargo, sensores más específicos como el giroscopio, acelerómetro o el sensor de una pantalla táctil están ligados a dispositivos como teléfonos móviles inteligente o tabletas. Cabe destacar que, algunos atributos dinámicos pueden ser generales, es decir, pueden ser recopilados independientemente del dispositivo utilizado, como las trazas de interacción con una interfaz gráfica, los nodos visitados y rutas seguidas para realizar una tarea concreta, etc.

A continuación, se definen una serie de perfiles para que las RPs puedan tener unas bases para seleccionar unos atributos u otros. Estos perfiles se basan en estudios del estado del arte que tratan de analizar la entropía o singularidad que proporcionan cada uno de estos atributos a la hora de generar una huella digital, así como la facilidad de recopilación de los mismos, y la posible aceptación que un usuario puede tener a la hora de permitir recopilarlos (debido a la invasión de la privacidad que puede suponer recopilar ciertos atributos). Para los atributos estáticos cabe destacar los trabajos [115], [116], mientras que para los atributos dinámicos el estudio realizado en [117], [118]. Estos perfiles distinguen, en primer lugar, por RPs accesibles desde web o desde un entorno móvil, y en segundo lugar se categorizan en tres niveles de seguridad (básico, medio y alto). Estos niveles de seguridad son incrementales, es decir, los niveles superiores incluyen los atributos seleccionados para los niveles inferiores. A continuación, se detallan los atributos específicos que se incluyen en cada uno de los niveles:

- Seguridad básica para una RP web: agente de usuario, lista de plugins, resolución de pantalla, zona horaria y habilitación de *cookies*.

- Seguridad básica para una RP en entorno móvil: Resolución de pantalla, zona horaria, utilización de batería y utilización del sistema.

- Seguridad media para una RP web: Canvas, WebGL y entorno.

- Seguridad media para una RP en entorno móvil: Canvas, giroscopio, acelerómetro y entorno.

- Seguridad alta para una RP web: dinámicas de ratón, dinámicas de teclado, interacciones con el entorno y geolocalización.

- Seguridad alta para una RP en entorno móvil: dinámicas de pulsación en pantalla, interacciones con el entorno y geolocalización.

Como se ha mencionado con anterioridad, estos perfiles son una recomendación, es decir, no son obligatorios y por consiguiente cualquier RP tiene la capacidad de partir de cualquiera de estos perfiles e ir añadiendo o eliminando atributos específicos para cumplir con sus requisitos y necesidades. Si una RP decide eliminar o no puede incluir alguno de los atributos específicos, para conseguir el mismo nivel de seguridad se recomienda seleccionar un atributo de un nivel de seguridad inmediatamente superior. Si esta casuística sucediese para el nivel más alto, la recomendación es seleccionar al menos dos atributos del nivel de seguridad inmediatamente inferior.

3.3.2. Generación de la huella digital

La información recopilada durante este paso ha de ser común a todos los usuarios y amplia en cuanto a volumen para cada usuario. Esto se debe a que, las huellas digitales han de poder ser comparadas entre los diferentes usuarios, por lo que, tienen que poseer las mismas características. Además, cuanta más información se recopile para cada usuario, se podrá determinar con mayor facilidad las fronteras de decisión y los patrones que definen a cada usuario de forma única. La tecnología utilizada ha de ser genérica y multiplataforma con el objetivo de que sea adaptativa. Los procesos de recopilación y generación de la información han de ser eficientes, evitando introducir latencias innecesarias y un consumo de recursos excesivos. La recopilación de los datos se puede llevar a cabo por lotes, es decir, recibiendo la información de forma periódica, o en tiempo real y de forma continua en el tiempo, es decir, en *streaming*.

Una vez se ha recopilado la información, el siguiente paso es el proceso de almacenamiento de la información. Este proceso tiene que hacer uso de un sistema fácilmente escalable, con el objetivo de que pueda aumentar sus capacidades en el caso de que el número de usuarios o de información recopilada sea masivo.

El procesado de los datos también debe estar claramente definido. Para ello, se deben descartar valores atípicos, información con ruido o información obsoleta, la cual puede hacer que los modelos generados se vean lastrados.

Es indispensable determinar como la información recopilada y posteriormente almacenada se va a representar. Por ejemplo, las cadenas de Markov son de

utilidad para transformar a estados cada comportamiento, acción o interacción del usuario. En este caso concreto, se analizarán las transiciones entre estados, determinando que, si el usuario se ha desplazado de un estado en el que es altamente probable que se encuentre, a un estado poco probable, el comportamiento realizado es atípico y por consiguiente se considera una anomalía y una posible brecha de seguridad. Otras soluciones tratan cada interacción del usuario como un conjunto de características que lo definen, por ejemplo, velocidades de movimiento o ángulo de desplazamiento en el caso de considerar dinámicas de movimientos de ratón. Estas características se agrupan temporalmente formando un vector que define el comportamiento de un usuario en un intervalo de tiempo.

En este trabajo se recomienda que una RP primero seleccione una solución simple que se adapte correctamente a sus necesidades. Esta solución puede basarse en generar secuencias de vectores, que se pueden agrupar posteriormente utilizando ventanas temporales deslizantes, de tal manera que los vectores consecutivos contengan también información de los vectores adyacentes. De esta forma, los vectores consideran una mayor cantidad de información de seguido. Por ejemplo, dado un vector que contenga cuatro elementos, estos se pueden separar en subvectores de longitud dos tal que, el primer subvector contiene los elementos uno y dos, el siguiente subvector los elementos dos y tres y así sucesivamente.

3.3.3. Modelado

En la parte de modelado, las técnicas de aprendizaje máquina basadas en detección de anomalías se posicionan como una de las mejores soluciones para detectar posibles brechas de seguridad utilizando información de comportamientos. Estas técnicas se denominan, en el ámbito del aprendizaje máquina, detección de atípicos. Además, estas técnicas permiten, entre otras cosas, que las RPs puedan manejar datos no etiquetados (aprendizaje no supervisado) y desequilibrados, los cuales son dos de los grandes problemáticas en este ámbito.

La detección de atípicos se basa fundamentalmente en establecer una medida de distancias o similitud (p. ej. la distancia Euclídea) entre objetos [85]. Para el caso concreto que aplica a esta tesis doctoral, estos objetos son huellas digitales de comportamiento. De esta forma, gracias a la medida de distancia o similitud, se pueden definir núcleos de comportamiento que se consideran normales o esperados para un usuario concreto teniendo en cuenta su huella digital, es decir, teniendo en cuenta su histórico de comportamientos. Posteriormente, las nuevas interacciones que llegan al sistema generan nuevas trazas de comportamientos que se comparan con el comportamiento esperado, es decir, con la huella digital. Definiendo un umbral,

se puede determinar si estos nuevos comportamientos son normales, o si por el contrario son atípicos y por lo tanto pueden suponer una brecha de seguridad.

Los algoritmos de aprendizaje máquina que pueden ser útiles para una RP con el objetivo de poder seguir el flujo de trabajo propuesto se pueden clasificar en tres grupos [119]:

- *Forecasting*: se basan en el aprendizaje supervisado. En este grupo, los algoritmos se entrenan utilizando datos etiquetados. Estos algoritmos tratan de realizar predicciones basándose en el análisis de tendencia, de estacionalidad y ciclos de datos temporales. ARIMA [120], SVM, K-NN, NN y las redes bayesianas son algunos ejemplos de algoritmos que encajan en este ámbito.

- No supervisado: se basan mayoritariamente en agrupar los datos. De esta forma, se toma como premisa que los datos generados por la misma fuente (es decir, generados por el mismo usuario) pertenecerán a los mismos grupos, mientras que los datos atípicos se encontrarán fuera de estos grupos definidos como normales. OC-SVM, K-means e *isolation forest* [121] son algunos ejemplos de algoritmos que encajan en este grupo.

- Basados en densidades: este grupo es en realidad una subcategoría de los no supervisados. Se basa en determinar la densidad de datos que contienen las distintas regiones del espacio. De esta forma, las regiones de alta densidad se pueden definir como las regiones normales o esperadas. Por otro lado, si una muestra se encuentra en una región de baja densidad, será considerado como atípica y, por consiguiente, se determina que no pertenece al usuario legitimo. Algunos ejemplos de estos algoritmos son *Local Outlier Factor* [122] oDBSCAN [123].

Una RP específica que desee implementar el flujo de trabajo propuesto, ha de primero analizar los datos recopilados y almacenados para generar la huella digital. Esta huella digital, en función de la naturaleza de la RP específica, poseerá unas cualidades únicas. Por ejemplo, una RP que recopile información de la sesión del usuario, podrá tener datos etiquetados y por lo tanto utilizar modelos de *forecasting* para realizar las predicciones o para mejorar las predicciones realizadas por modelos no supervisados. Cabe destacar, que la información de las etiquetas también puede estar sesgada, es decir, si durante el proceso de recopilación de datos ya existía una brecha de seguridad, los datos recopilados pueden contener comportamientos atípicos. Esto significa, que estos datos atípicos serán considerados como normales o esperados si no se realiza un proceso de preprocesamiento y limpieza de datos adecuado. La recomendación que se propone en esta tesis doctoral es utilizar las etiquetas siempre y cuando

sea posible, pero teniendo en cuenta el propio sesgo que puede existir en dichos datos y, por consiguiente, combinar modelos supervisados con modelos no supervisados o basados en densidades para obtener predicciones más precisas. Por ejemplo, una buena aproximación de combinación de varios tipos de algoritmos se encuentra en [124], donde primero se aplican técnicas de *clustering* para agrupar los diferentes comportamientos para posteriormente aplicar múltiples algoritmos específicos que modelen cada grupo de forma independiente, obteniendo resultados más precisos para cada uno de ellos.

Todos los tipos de algoritmos, arriba mencionados, se pueden implementar de dos formas diferentes, *offline* y *online*. Los algoritmos *offline* utilizan el histórico de información para generar un modelo estático que posteriormente se utiliza para realizar predicciones durante un periodo de tiempo. Estos modelos han de irse adaptando para poder considerar los cambios de la distribución de los datos de entrada, es decir, los usuarios no mantienen su comportamiento estable a lo largo del tiempo. Para poder considerar estos cambios de comportamiento, estos modelos han de irse reentrenando. Para lograr que el reentrenamiento sea efectivo, se deben establecer unas políticas que contemplen indicadores para determinar cuándo las distribuciones iniciales de los datos han cambiado y por lo tanto el modelo ha quedado obsoleto y está perdiendo eficacia y precisión.

Por otro lado, los algoritmos *online* van entrenando a medida que van realizando nuevas predicciones. De esta forma, cuando llegan nuevas muestras para evaluar, el propio modelo se va actualizando y cambiando sus fronteras de decisión para considerar estos nuevos comportamientos. Así, a medida que el modelo obtiene nuevas muestras, su eficacia y precisión va aumentando. Sin embargo, también hay que considerar que estos modelos también han de ir descartando la información obsoleta que puede quedar cuando el modelo lleva funcionando durante un largo periodo de tiempo. Este tipo de modelos son de mucha utilidad cuando se combinan con una recopilación de la información en *streaming*.

Una de las problemáticas más comunes a las que se enfrentan las propuestas de este ámbito, es la de detectar anomalías para los primeros accesos, es decir, cuando apenas se posee información de comportamiento del usuario. Esta problemática se denomina arranque en frío. Para este caso concreto, el resultado suelen ser modelos de aprendizaje máquina muy pobres, que no poseen suficiente información para comparar las muestras a evaluar y, por consiguiente, categorizan multitud de muestras como anomalías, cuando no lo son. Esto se traduce, en modelos con una tasa de falsos positivos muy elevada, haciendo que los sistemas en los que se integran estos modelos no sean muy fiables. Una buena solución para este problema es utilizar un sistema de recomendación, capaz de detectar cuando un usuario está utilizando

Figura 3.4. Matriz de confusión

		Predicción	
		Positivo	Negativo
Valor Real	Positivo	VP	FP
	Negativo	FN	VN

el sistema por primera vez, y alimentar una máquina de factorización para tratar de asemejar estos comportamientos de los de otros usuarios parecidos, mitigando así el problema [125]. Otras posibles soluciones se basan en utilizar modelos no paramétricos para la fase inicial en la que puede existir arranque en frío y otros modelos más precisos cuando ya se ha superado esta fase [126], o tomar como punto de partida un modelo previamente entrenado para otro usuario. De esta forma, aunque siga existiendo el arranque en frío, este durará un menor tiempo, pues el modelo no parte de cero y solo tendrá que irse adaptando, modificando sus fronteras de decisión, a los nuevos comportamientos que genere el nuevo usuario.

3.3.4. Evaluación

El paso de evaluación permite cuantificar la eficacia de los modelos de aprendizaje máquina desarrollados en el paso anterior. De esta forma, se pueden encontrar errores de diseño, ya sea a la hora de seleccionar o generar la huella digital, como determinar los motivos que hacen que el modelo en sí no esté funcionando correctamente. Esta detección temprana permite que la RP pueda regresar a los pasos anteriores, con el objetivo de mejorar la eficacia del modelo final. Es por esto que este paso es fundamental, pues se posiciona como la capa intermedia entre el desarrollo del flujo de trabajo y su puesta en producción.

En primer lugar, se definen las clases que se quieren evaluar. Para el caso del análisis de comportamiento, típicamente se consideran dos clases, la positiva que representa al usuario genuino, y la clase negativa que representa al usuario impostor. Dadas estas dos clases, un modelo de aprendizaje máquina puede predecir de forma correcta o incorrecta cada una de estas clases, obteniendo así cuatro posibles valores. En primer lugar, el valor Verdaderos Positivos (VP) representa a los usuarios genuinos autenticados correctamente. Verdaderos Negativos (VN) representa a los impostores rechazados por el modelo de forma correcta. Falsos Positivos (FP) son los usuarios impostores que se han logrado autenticar como usuarios genuinos y no han sido detec-

tados por el modelo. Falsos Negativos (FN) son los usuarios genuinos que se han detectado incorrectamente como impostores por el modelo. Estos cuatro valores, se agrupan formando una matriz de confusión tal y como se muestra a continuación (ver Figura 3.4):

Existen multitud de métricas para evaluar diferentes aspectos específicos de los resultados obtenidos para los algoritmos de aprendizaje máquina. Dada esta matriz de confusión, típicamente se suelen calcular tres métricas asociadas: la exactitud, la precisión y la sensibilidad. La exactitud es la tasa de acierto sobre el total de muestras, es decir, el porcentaje de acierto del modelo de manera global. La precisión es el porcentaje de aciertos para la clase positiva. La sensibilidad es el porcentaje de muestras para la clase positiva que se aciertan de forma correcta sobre el total de posibles positivos. Estas métricas vienen definidas como sigue:

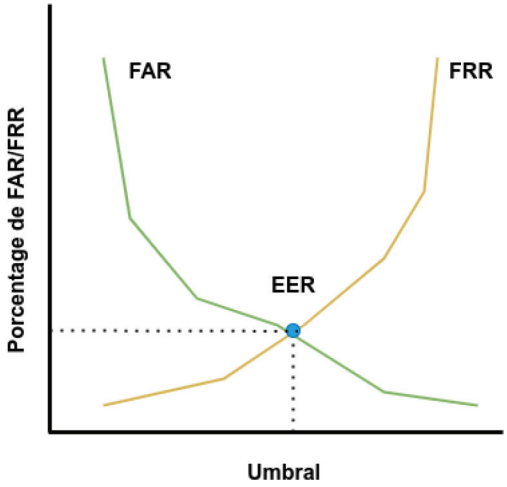
- Exactitud= (VP+VN)/(VP+FP+FN+VN)
- Precisión= VP/(VP+FP)
- Sensibilidad= VP/(VP+FN)

Sin embargo, en el ámbito del modelado de comportamiento, estas métricas no suelen ser suficiente, pues los problemas de este tipo poseen unas cualidades peculiares, como los datos desequilibrados. Por ejemplo, en el caso de considerar una clase mayoritaria que posee el 99 % de la información genuina, es decir, de información de clase positiva, un modelo trivial que clasifique cada muestra en esta clase obtendría un 99 % de tasa de acierto o exactitud. Si una RP únicamente tuviese en cuenta esta métrica, seguramente este modelo sería seleccionado, pues tiene un acierto muy alto. Sin embargo, el objetivo del problema es justo detectar la clase minoritaria, es decir, la clase que contiene los comportamientos realizados por usuarios impostores (clase negativa) y que pueden suponer una brecha de seguridad. En este caso, este modelo trivial tendría una tasa de acierto del 0 %, convirtiendo así a este modelo en inútil y por tanto habría que desecharlo.

Una métrica que ha demostrado ser más efectiva que las anteriores para lidiar con datos desequilibrados es el F1-score [127]. Esta métrica se puede definir como la media armónica entre la precisión y la sensibilidad. Nuevamente, esta métrica considera como clase de interés la clase genuina.

Todas las métricas definidas hasta ahora se basan fundamentalmente en evaluar la clase positiva, es decir, la de usuarios genuinos. Para el problema que se desea resolver en esta tesis, debería ser al contrario, es decir, se debe evaluar la clase de usuarios impostores, pues es la clase de interés.

Figura 3.5. Representación del Equal Error rate (EER) en función del False Acceptance Rate (FAR) y el False Rejection Rate (FRR)



Existen métricas específicas que detallan con más precisión las casuísticas específicas en el ámbito del análisis de comportamiento. Estas métricas son: tasa de aceptación falsa (en inglés, *False Acceptance Rate* [FAR]), tasa de rechazo falso (en inglés, *False Rejection Rate* [FRR]) y la tasa de error igual (en inglés, *Equal Error Rate* [EER]) [128]. El FAR se define como la proporción de usuarios impostores que no son detectados correctamente por el modelo. El FRR es la proporción de usuarios legítimos que el modelo predice como usuarios impostores. Un mismo modelo puede ser más restrictivo o más permisivo en función del umbral definido. Esto es, un mismo modelo puede determinar la probabilidad de una muestra de pertenecer a la clase genuina o a la clase impostora, pero es gracias al umbral cuando finalmente se clasifica cada muestra. De esta forma, un mismo modelo con un umbral alto puede ser más restrictivo y por lo tanto tener más acierto para la clase impostora pero menos para la clase genuina (sistema más seguro), o puede ser más permisivo fijando un umbral bajo y por consiguiente tener más acierto en la clase genuina pero menos en la impostora (nunca se podría fijar un umbral que haga tener más acierto en ambas clases simultáneamente). De este modo, dependiendo del umbral se obtienen una pareja de valores FAR y FRR. El EER se puede definir, por tanto, como el punto óptimo entre estos dos valores, es decir, el umbral donde se obtiene un mejor valor para ambas métricas de forma simultánea (ver Figura 3.5).

Las métricas utilizadas en este documento para evaluar los modelos y que se recomienda que utilice cualquier RP que desee integrar el flujo de trabajo propuesto, se definen a continuación.

- Especificidad= VN/(VN+FP)
- FAR= FP/(FP + VN)
- FRR= FN/(FN + VP)
- EER= Valor de la intersección entre FAR y FRR cuando se consideran múltiples umbrales.
- Valor Predictivo Negativo (VPN)= VN/(VN+FN)
- $F1^{-} = 2 \cdot \frac{\text{VPN} \cdot \text{Especificidad}}{\text{VPN} + \text{Especificidad}}$

Donde la especificidad es la proporción de impostores correctamente identificados, esto es, la sensibilidad para el grupo de impostores (clase negativa). VPN es la eficacia del método a la hora de detectar impostores, esto es, la precisión para el grupo de los impostores. $F1^{-}$ es la media harmónica entre el VPN y la especificidad [129], es decir, el F1-Score para el grupo de los impostores.

3.3.5. Integración en los sistemas de gestión de identidades federados

La integración del flujo de trabajo en los esquemas de gestión de identidades federados acarrea multitud de retos.

En primer lugar, los usuarios finales seguramente no estén dispuestos a tener que instalar software externo, especialmente si este va a afectar a su privacidad. Este flujo de trabajo es claramente un ejemplo en el que los usuarios pueden sacrificar cierta privacidad (dependiendo de la huella digital seleccionada) con el objetivo de ganar seguridad. Por ejemplo, un usuario puede ser reacio a que una red social recopile información de su comportamiento, sin embargo, si su aplicación de banca electrónica se lo solicita puede optar por aceptar las condiciones.

El flujo de trabajo aquí expuesto no requiere que el usuario final tenga que instalar software externo, simplemente necesita que el usuario acepte las condiciones y por consiguiente consienta que la RP recopile información para generar su huella digital de comportamiento. Es por esto que un mismo usuario puede optar por permitir que ciertas RPs recopilen más información que otras, o incluso ninguna, con el objetivo de que el propio usuario final se vea beneficiado o no, de la implantación del flujo de trabajo propuesto. En cuanto

a la privacidad, la RP ha de informar de forma transparente de todos y cada uno de los datos que va a recopilar, así como de cuál va a ser el uso que le va a dar. Cabe destacar, que esta información solo ha de ser accessible por la propia RP, es decir, no se tiene que compartir con terceros, y ha de ser debidamente protegida, desde que se genera en el dispositivo del usuario, mientras se envía por algún canal de comunicación debidamente encriptado utilizando *Transport Layer Security* (TLS), hasta que se almacena en algún sistema de información.

Por otro lado, la aceptación de los usuarios también se va a ver afectada dependiendo de la eficiencia y la usabilidad. La solución finalmente implementada ha de estar basada en una huella digital lo suficientemente única y robusta como para poder detectar las brechas de seguridad, sin afectar altamente al consumo de recursos, evitando introducir latencias innecesarias y disminuyendo el número de falsos positivos que pueden hacer que el sistema se vuelva poco usable.

Los IdPs no se deben ver afectados por la implantación de este flujo de trabajo. Todos los aspectos relacionados con la integración (consumo de recursos, notificación a los EU, etc) han de recaer sobre la propia RP, la cual es la interesada en proporcionar a sus usuarios finales la capacidad de aumentar sus niveles de seguridad. Es por esto que la implantación de este flujo de trabajo solo dependerá de la comunicación entre la RP y el EU, no necesitando de la colaboración de terceros o del IdP.

Otro reto asociado a la integración del flujo de trabajo es la aceptación de las propias RPs. Esto se debe a que, si el flujo de trabajo requiere modificar los propios estándares de gestión de identidades existentes, probablemente se van a volver reacias a implementarlos. Esto se debe a la dificultad añadida de implantación que esto supondría. Es por esto que los flujos de información de los estándares federados no han de ser modificados, o en su defecto lo menos posible y, por lo tanto, las RPs deben de utilizar los propios mecanismos facilitados por dichos estándares para realizar su cometido. De este modo, se deben utilizar las propias peticiones y respuestas, los parámetros, *tokens* y cualquier otro aspecto ya existente en los flujos de información para realizar la implantación del flujo de trabajo.

3.3.6. Resumen del flujo de trabajo

A modo resumen, en la Tabla 3.1 se muestra de forma gráfica, todas las decisiones y consideraciones que ha de tomar una RP para implantar el flujo de trabajo propuesto. Estas decisiones se corresponden con cada uno de los pasos fundamentales en los que se basa dicho flujo de trabajo.

Tabla 3.1. Resumen del flujo de trabajo propuesto

Tarea	Decisión	Alternativas	Criterio de decisión
1. Selección de huella digital	Huella digital que proporciona la suficiente singularidad y no es invasiva para el usuario	Atributos estáticos y dinámicos (Figure 3.3)	Caso de uso. Niveles de seguridad deseados.
2. Generación de la huella digital	Recopilar, almacenar, preprocesar y representar los datos: tecnologías y procedimientos	Por lotes o streaming; SQL o No-SQL; lenguajes de programación; Técnicas de representación de la información	Eficiencia y eficacia, Consumo de recursos, escalabilidad
3. Modelado	Detectar anomalías en las huellas digitales: técnicas	*Forecasting*, aprendizaje no supervisado, algoritmos basados en densidades; modelos *offline* o *online*	Huella digital seleccionada, caso de uso, recursos compu- tacionales disponibles en la RP
4. Evaluación	Decidir si se necesitan más iteraciones en el flujo de trabajo: métricas de evaluación	Exactitud, especificidad, FAR FRR, EER, VPN y $F1^{-}$	Caso de uso
5. Integración	La RP debe integrar los modelos de análisis de comportamiento para realizar los procesos de IAAA: modificaciones	Cambios en el estándar (peticiones, *tokens*, flujos); Cambios en la implementación (comprobaciones adicionales, estructura de datos)	Caso de uso, coste permitido