Universidad de Lima

Facultad de Ingeniería

Carrera de Ingeniería de Sistemas

METODOLOGÍA PARA LA GENERACIÓN DE UN DATASET DE DINÁMICA DE TECLEO BASADO EN UN ENTORNO WEB

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Aron Lo Li

Código 20160795

Asesor

Juan Manuel Gutierrez Cardenas

Lima – Perú

Junio de 2024

Metodología para la generación de un dataset de dinámica de tecleo basado en un entorno web

Lo Li, Aron
20160795@aloe.ulima.edu.pe
Universidad de Lima

Resumen:

Los sistemas de autenticación basados en dinámica de tecleo identifican a sus usuarios por medio del análisis del patrón habitual de tecleo cuando estos interactúan con un dispositivo de entrada que podría ser el teclado de una computadora. Dentro de la literatura, existen varias investigaciones que mencionan el uso de distintas técnicas como los modelos de distancia o incluso el aprendizaje de máquina. Estos en conjunto con las diferentes agrupaciones de atributos de tecleo como pueden ser el tiempo de presión de la tecla, el tiempo en que se mantiene en el aire, la latencia entre las pulsaciones; entre otras, permiten la identificación de las identidades de los usuarios. Sin embargo, de los varios trabajos revisados, existen muchos que emplean datasets generados por los propios investigadores que además de ser privados, carecen de la documentación del cómo estos fueron recopilados y creados.

Esto representa un gran problema para los futuros investigadores que buscan mejorar o avanzar con los trabajos previos, ya que la posibilidad de replicar estos trabajos es casi nula debido a la falta de documentación además del no poder contar con los datasets usados en dichos papers respectivamente. Asimismo, al utilizar propias bases de datos, los investigadores no ofrecen una buena visión general del rendimiento global de sus métodos, sino una visión general de un caso específico: el representado por su base de datos. En esta investigación, se planteó una metodología para el desarrollo de un dataset público cuya recolección de los datos se realizó por medio de una herramienta desplegada en el navegador, donde los usuarios participaban remotamente desde sus computadoras personales. Además, se aseguró que durante las recopilaciones, se cumplieran con protocolos de recolección consideradas como buenas prácticas por diversos investigadores. Esto permite a los futuros investigadores el tener la posibilidad de contar con un dataset de buena calidad y que refleje de manera más realista el tecleo de los usuarios.

Palabras Clave: Seguridad de la Computación, Aprendizaje de Máquina, Dinámica de Tecleo, Verificación de Usuario, Conjunto de Datos.

Abstract:

Keystroke dynamics-based authentication systems identify their users by analyzing the typical keystroke pattern when they interact with an input device that could be a computer keyboard. Within the literature, there are several investigations that mention the use of different techniques such as distance models or even machine learning. These in conjunction with different groupings of characteristics or typing traits such as flight time, pressure time, typing latency, among others, they allow the identification of user identities. However, among the various works reviewed, there are many that use data sets generated by the researchers themselves that, in addition to being private, lack documentation of how they were collected and created.

This represents a great problem for future researchers who seek to improve or advance with previous work, since the possibility of replicating these works is nearly non-existent due to the lack of documentation in addition to not being able to count on the datasets used in their papers respectively. Furthermore, by using their own databases, researchers do not provide a good overview of the global performance of their methods, but rather an overview in a specific case: the one represented by their database. In this research, a methodology was proposed for the development of a public dataset whose data collection was carried out through a tool displayed in the browser, where users can participate remotely from their personal computers. In addition, it was ensured that certain protocols, such as those mentioned above, are complied with during the collections. This will allow future researchers to have the possibility of having a good quality dataset that more realistically reflects the typing of users.

Keywords: Computer Security, Machine Learning, Keystroke Dynamics, User Verification, Dataset

1. INTRODUCCIÓN

Los ataques de robo de identidad representan uno de los principales desafíos en materia de ciberseguridad para las grandes empresas en la actualidad. De acuerdo con Thomas (2018), se estima que casi un 80% de las grandes corporaciones del mundo han experimentado al menos uno de estos tipos de ataques en los últimos años. Estos ataques se han vuelto tan recurrentes que se pueden ver noticias de ataques hacia empresas prominentes como JP Morgan, Target o Sony. Un ejemplo reciente de estos ataques, dirigido hacia este último, según cifras oficiales, generaron pérdidas que alcanzaron los 171 millones de dólares en el 2016 (Lisa, 2014). La implementación de técnicas de autenticación sólidas, como las biométricas, puede contribuir a reducir estos costos al garantizar que solo los usuarios autorizados puedan acceder a los sistemas correspondientes.

En la actualidad, se emplean diversas técnicas para validar la identidad de las personas, siendo consideradas las biométricas entre las más fiables. Esta se refiere al uso de características humanas que hacen que cada individuo sea único y abarcan cualquier característica personal que pueda utilizarse para verificar de forma exclusiva la identidad de una persona (Dee, et al., 2019). De acuerdo con Darabseh y Namin (2016), estas se dividen en dos grandes categorías: las características biométricas físicas, como el reconocimiento de la huella dactilar, el facial o del iris; y las características conductuales, que incluyen a modo de ejemplo, la identificación del patrón del caminar de una persona (*gait*), las firmas manuscritas (*handwritten signature*) o la dinámica de tecleo (*keystroke dynamics*).

Según Monrose y Rubin (2000), la dinámica de tecleo implica analizar el patrón de tecleo habitual de un usuario cuando este interactúa con un dispositivo de entrada, como el teclado de una computadora. Varias investigaciones, como las llevadas a cabo por Baynath, et al. (2018), Darabseh (2016) y Çeker, et al. (2015), emplean el uso de distintas técnicas como los modelos de distancia o incluso el aprendizaje de máquina. Estas en combinación con diferentes agrupaciones de atributos de tecleo como pueden ser el tiempo de presión de la tecla, el tiempo en que se mantiene en el aire, la latencia entre las pulsaciones; entre otras, permiten la identificación de las identidades de los usuarios.

El uso de la dinámica de tecleo ofrece varias ventajas en comparación con otros métodos de autenticación: En primer lugar, es práctico y computacionalmente viable, además de la facilidad que existe de capturar estos tipos de datos para alimentar los modelos de clasificación. Además, está forma de autenticación es económica, ya que no necesita un dispositivo de hardware especial, sino que solo requiere el teclado del mismo ordenador. Por último, los patrones de escritura pueden permanecer disponibles después de que la fase de autenticación haya concluido, lo cual es una representa una gran ventaja.

Distintas investigaciones (Baynath, et al., 2018; Chen, et al., 2007; Çeker & Upadhyaya, 2015; Darabseh & Namin, 2016; Kang et al., 2007; Zhong & Deng, 2015; Zhong, et al., 2012) mencionan el empleo de distintas técnicas, en combinación con diferentes agrupaciones rasgos de tecleo como pueden ser el tiempo de vuelo, el tiempo de presión, la latencia de tecleo, entre otras para identificar las id de los usuarios, dando resultados alentadores y de una precisión aceptable. No obstante, de los varios trabajos revisados en la literatura, existen muchos que emplean datasets generados por los propios investigadores que además de ser privados, carecen de la documentación del cómo estos fueron recopilados y creados. También se debe de mencionar que un factor casi común en los trabajos revisados anteriormente es que el registro del tecleo de los usuarios se dieron en entornos de laboratorio controlados, donde se aseguró que los usuarios tuvieran un "previo entrenamiento" en la forma de teclear de cierta frase. Esto se puede ver por ejemplo en el trabajo del dataset de GREYC (2009), donde todos los participantes usaron la misma laptop para realizar sus pruebas dentro de una sala acondicionada del laboratorio, además de que se les permitió tener una sesión inicial para que se puedan familiarizarse. Este procedimiento podría generar cierto sesgo en el tecleo de los usuarios, pues afecta de manera inconsciente en la forma de teclear habitual de las personas. En otros estudios como el de Chen (2007), se contó con la participación 100 personas y se les solicitó que tecleen la palabra "try4-mbs". Si bien hubo una gran cantidad de participantes, las capturas de estas muestras se dieron en pocas sesiones y no se especifica en mayor detalle los protocolos que se siguieron para garantizar la calidad del dataset, que son necesarios para el entrenamiento de estos tipos de modelos de dinámica de tecleo.

Hablando de manera general, esto representa un gran problema para los futuros investigadores que buscan mejorar o avanzar con los trabajos previos, ya que la posibilidad de replicar estos trabajos es casi nula debido a la falta de documentación además del no poder contar con los datasets usados en dichos artículos respectivamente.

En este sentido, se plantea una metodología para el desarrollo de un dataset público cuya recolección de los datos se hará por medio de una herramienta desplegada en el navegador, donde los usuarios podrán participar remotamente desde sus computadoras personales. Además, se asegurará que durante las recopilaciones se tomen en consideración diversas recomendaciones para garantizar la calidad de los registros. Por ejemplo, Giot, et al. (2009) menciona que "… para crear una buena base de datos biométrica conductual, el número de sesiones requeridas debe de ser superior o igual a tres, que estas sesiones deben estar espaciadas en el tiempo, la población debe ser grande y diversificada. Asimismo, al utilizar propias bases de datos, los investigadores no ofrecen una buena visión general del

rendimiento global de sus métodos, sino una visión general en un caso específico: el representado por su base de datos." Seguir estas buenas prácticas permitirá a los futuros investigadores el tener la posibilidad de contar con un dataset de buena calidad y que refleje de manera más realista el tecleo de los usuarios. El resto del documento está organizado de la siguiente manera: la sección 2 hace una descripción del estado del arte y de la literatura. En la sección 3, se encuentran los antecedentes, donde se explicarán conceptos importantes para entender el trabajo de investigación. En la sección 4 se encuentra la metodología a seguir junto a la prueba de concepto, mientras que en la parte 5 se muestran los resultados de los experimentos realizados. En la parte 6 encontraremos la discusión de resultados, donde después de esto, finalizamos con la sección de conclusiones.

2. ESTADO DEL ARTE

Esta sección presenta una recopilación de los distintos datasets y protocolos de recolección empleados en los trabajos de clasificación para la autenticación del usuario usando dinámica de tecleo.

Existen distintas investigaciones que mencionan la creación de datasets propios. Por ejemplo, en el trabajo de Bleha, et al. (1990), se propuso la recopilación de los registros del tecleo de 26 usuarios por medio de un programa implementado en Fortran e instalado en una computadora personal de la marca IBM durante 8 semanas. Los voluntarios tenían que escribir tanto su nombre personal, así como una frase elegida por los investigadores como contraseñas. El programa se encargaba de capturar y calcular el intervalo de tiempo que tomaba presionar cada tecla. En cuanto a la frecuencia de recolección, los usuarios eran libres de escoger el momento en que querían participar basados en su disponibilidad. En el trabajo de Aráujo, et al. (2005), se recolectaron los registros de tecleo de 30 participantes. Estos tenían que escribir una palabra fija impuesta de al menos 10 caracteres. Entre los datos que se capturaron en el experimento, se almacenaron los códigos de las teclas, así como el timestamp del momento en que se pulsaba y se soltaba cada tecla. En cada sesión de recopilación, si bien no se menciona cuando duraba, se indica que se recolectaron alrededor de 10 muestras de la palabra por cada usuario. Hosseinzadeh, et al. (2008) propuso la recolección de las muestras de tecleo de 41 usuarios durante 4 semanas, de los cuales 30 eran hombres y 11 eran mujeres. La edad promedio entre los usuarios era de 30. Para cada usuario, se le dio una aplicación especialmente diseñada para la recolección de patrones de tecleo al cual le denominaron KbApp. Este capturaba el código de la tecla presionada y los timetamps del momento de la presión y liberación de cada key. Se almacenaron 1153 registros promediando una muestra hecha por cada usuario en cada día y cada sesión le tomaba al participante alrededor de 5 minutos. En la investigación de Çeker, et al. (2015), se registraron las muestras de 39 voluntarios en varias sesiones que duraron aproximadamente 11 meses. En cada semana, se hacían dos tipos de sesiones diferentes, los cuales tenían una separación de al menos 2 días. La primera sesión consistía en la resolución de un cuestionario, donde se registraba el tecleo de las respuestas de cada pregunta. La segunda sesión consistía en transcribir un párrafo fijo procedente de una conferencia dada por Steve Jobs en la universidad de Standford. Cada sesión duraba aproximadamente una hora. Una característica casi común en estos trabajos es el uso de entornos controlados para registrar los patrones de tecleo de los usuarios, pues los investigadores definen la palabra que se usará para el experimento y se les pide a los usuarios ingresar sus muestras de tecleo en una computadora que se les proporcionada para las pruebas.

Cheng, et al. (2007) propone la creación de un dataset público para su uso en técnicas de dinámica de tecleo llamado "Keystroke 100 Dataset". Se contó con la participación de 100 usuarios voluntarios, los cuales tuvieron que escribir 10 veces la palabra "try4-mbs". Se capturaron la latencia del tecleo, así como la fuerza de presión, en unidades de tiempo de milisegundos. Entre los protocolos, se les pidió a los participantes que se familiarizaran con la palabra antes de comenzar a grabar los registros de cada sesión. No se menciona el periodo de la duración del experimento, ni el tiempo promedio que tomaba las sesiones de los usuarios.

El "CMU Keystroke Dynamics - Benchmark DataSet" se desprende de la investigación elaborada en el trabajo de Killourhy, et al. (2009). Se registraron aproximadamente 400 muestras de tecleo de 51 usuarios. Estos tuvieron que escribir 50 veces la palabra ".tie5Roanl" en cada sesión, de las cuales hubieron ocho en total. En cada registro se capturaron los siguiente features: "holdtime", "keyDownDown" y "KeyUpDown", los cuales todos están almacenados en la unidad de tiempo de segundos (número de tipo punto flotante). Las sesiones duraban alrededor de 3 minutos en promedio y se hacían en al menos un día entre sesiones para capturar la variación del tecleo de los distintos usuarios. La captura de la palabra tenía que ser tipeada de manera correcta y sin opción a corrección usando la tecla "*backspace*", caso contrario se descartaba la muestra y se solicitaba otra nuevamente.

El dataset GREYC propuesto en el trabajo de Giot, et al. (2009) realiza una recopilación de los patrones de tecleo de 133 participantes. Estos tenían que teclear la palabra fija "greyc laboratory" por al menos dos meses. Los usuarios tenían que participar en la sesión de cada semana, aunque algunos hacían hasta dos. En cada sesión se capturaban las latencias de tecleo, las cuales se almacenaban en un archivo de "Sqlite" en la unidad de tiempo de "ticks". Se almacenaron un total de 7555 registros teniendo un promedio de 51 muestras por usuario. No se menciona el tiempo aproximado que duraban las sesiones, pero se indica que entre las cantidades que se tienen recolectados por usuario, la mayoría presenta al menos más de cinco sesiones.

3. ANTECEDENTES

3.1 Autenticación e identificación del usuario

Existen dos conceptos que, aunque parezcan similares, presentan enfoques distintos para lograr sus objetivos; los cuales son: la autenticación y la identificación. La autenticación del usuario, como lo mencionan en el trabajo de Abramson M. (2013), es definida como el proceso de algún usuario reclamando ser una persona específica. La autenticación requiere que el reconocimiento sea un problema de relación de uno vs uno, donde el sistema trata de validar que la entrada de un usuario corresponda con la ya previamente guardada de este mismo (el cual es a lo que se enfocará este trabajo), mientras que la identificación es más un problema de uno vs varios, donde una entrada va a ser analizada contra todos los registros almacenados de los usuarios, obteniendo al final el perfil con el cual tiene mayor similitud. Monrose F. (2000) menciona que el problema de la autenticación "… se puede clasificar en tres grandes grupos: aquellos que la persona conoce como una contraseña, algo que la persona posee como una tarjeta de identidad o un token electrónico y el tercer grupo, el cual es usar las características de la persona, donde este último se conoce como biometría."

3.2 Biometría

De acuerdo con Dee (2019), la biometría implica la utilización de atributos distintivos de cada ser humano y abarca cualquier rasgo personal que pueda emplearse para verificar de manera única la identidad de un individuo. Tal como se señala en el artículo de Namin (2016), entre las distintas formas de biometrías, se pueden reconocer dos categorías principales: las características biométricas físicas, como el reconocimiento de la huella dactilar, facial o del iris; y las características conductuales, donde se encuentran los rasgos biométricos como la identificación del patrón de caminar de una persona (*gait*), las firmas manuscritas (*handwritten signature*) o la dinámica de tecleo (*keystroke dynamics*), el cual este último se explicará con más detalle a continuación.

3.3 Dinámica de tecleo

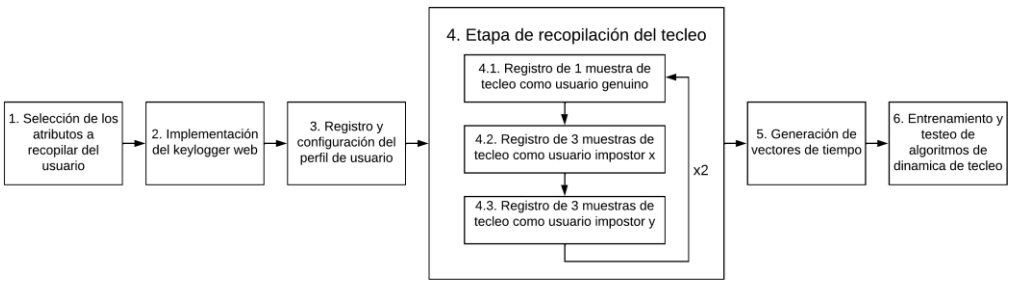
La dinámica de tecleo es definida por Darabseh (2016) como "… una característica biométrica conductual que implica analizar el patrón de escritura habitual de un usuario cuando interactúa con el teclado de una computadora". Para el diseño de un sistema como este, Monrose (2000) establece que es necesario considerar tres etapas claves: la representación, la extracción y la clasificación. La representación de los datos de entrada mide las características del patrón del objeto a reconocer (identidad), que en este caso sería los features. Un ejemplo sería los tiempos de presión de las teclas cuando el usuario está usando el computador. Estas generalmente son almacenadas en unidades de milisegundos en forma de un vector de n-dimensiones, donde n es la cantidad de features o características consideradas para la representación de la identidad. La segunda etapa es la de extracción de las características de los datos de entrada, que básicamente es medir y obtener los valores de los atributos establecidos en la etapa anterior por cada usuario. Generalmente en esta etapa se suele acompañar de técnicas como la reducción de dimensionalidad, que permite como su nombre lo dice, reducir el número de variables que serán usadas en la siguiente etapa. A menudo, a esto último se le suele denominar problemas de preprocesamiento y extracción de características. La última etapa es la clasificación e identificación que implica la determinación de procedimientos de decisión óptimas, que después de que los datos observados de los patrones a ser reconocidos se hayan expresado en forma de vectores de medición, se decidirá a qué clase de patrón pertenecen estos datos.

4. METODOLOGÍA

En esta sección, se presenta la metodología usada para la creación del dataset. Esta consta de distintas etapas que se pueden ver de manera resumida en la Figura 1.

Figura 1

Diagrama de bloques de la metodología propuesta



La primera etapa hace referencia a la selección de los atributos a recopilar del usuario, los cuales pueden ser tanto cualitativos como cuantitativos. En cuanto a los cualitativos, se recolectaron datos personales como: el nombre de usuario, la contraseña, el género, la mano dominante y el consumo de algún medicamento y/o enfermedad que pudieron afectar con la motricidad de las manos durante las pruebas. Estos fueron capturados en el momento en que los usuarios ingresaban por primera vez a la plataforma web y se les pedía completar un formulario. En lo referente a los datos cuantitativos, se recolectaron los datos como: la edad, la hora en que se realizó la recolección de la muestra en formato de "timestamps". También se recopiló el código de cada tecla presionada, para los eventos de "press" y "release", donde se le asoció la marca de tiempo de la computadora, la cual fue almacenada en la unidad de tiempo de milisegundos. Adicionalmente, se capturaron tanto la dirección IP así como el "user client" del navegador, que podrían ser un dato importante a la hora de analizar el contexto en el que se realizaron las pruebas.

En el segundo punto, se menciona la implementación de la herramienta que servirá de keylogger. Se optó por la creación de una página web, la cual tenía principalmente dos campos de texto, donde el usuario tenía que teclear tanto su nombre de usuario como la contraseña. Cualquier palabra que se ingresaba en esos cuadros de texto activaba un script hecho en javascript que capturaba esos datos y los almacenaba localmente. Al finalizar la sesión, el usuario tenía que oprimir un botón que enviaba los registros a un servidor desplegado en la nube. Una de las ventajas de este enfoque web, es que permitía a los participantes acceder a la herramienta desde su computador personal, donde realizaban las pruebas en un entorno más familiarizado. Esto permitió obtener datos simulando un escenario más realista.

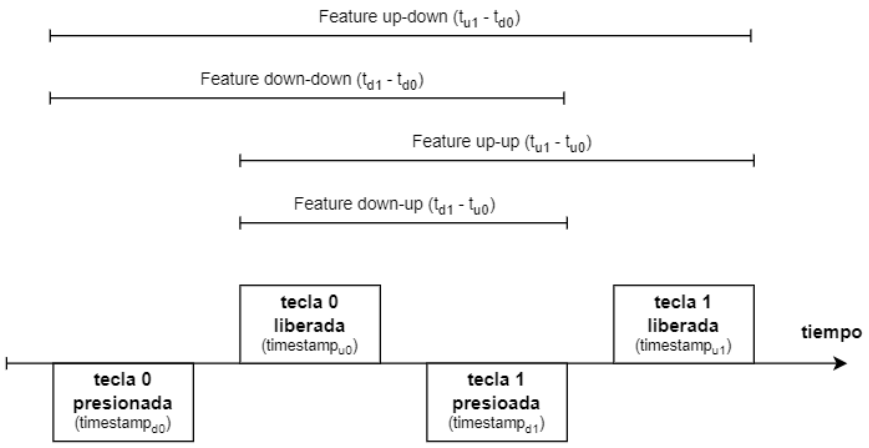
En la tercera etapa, el usuario que ingresaba por primera vez a la herramienta debía de registrar los datos personales que se le solicitaba y definir en el sistema el nombre de usuario y la contraseña que tendría que usar para todas las sesiones de recopilación. Con respecto al nombre del usuario, si bien no se aplicó una restricción de la palabra que podrían escoger, se les recomendó que esta sea una palabra simple, que no contengan de preferencia caracteres especiales y que tenga al menos una longitud de nueve letras. Estas sugerencias también son mencionadas en el artículo de Killourhy (2009) y tienen como objetivo el permitir al usuario escoger una palabra que le genera la sensación de comodidad y naturalidad al escribir. En cuanto a la referente a la contraseña, para dar una mayor variabilidad al experimento, los participantes fueron divididos en dos grupos, aquellos que tenían que ingresar al sistema con una contraseña de longitud fija impuesta por el investigador, mientras que el otro grupo podía ingresar una contraseña de elección libre y sin restricciones. Esta población fue segmentada en una proporción de 70:30 respectivamente, los cuales fueron seleccionados de manera aleatoria por la misma plataforma durante la fase de inscripción. Esto nos permitió tener una proporción de muestras más variadas que podrán ser usadas en posteriores investigaciones, dado que no se han encontrado en la literatura datasets cuyas contraseñas hayan sido elegidas por el propio usuario. Con respecto a la proporción elegida, se tomó en consideración la cantidad de participantes de los trabajos mencionados en la literatura [26, 30, 41, 39, 51, 100] y se sacó un promedio simple de los participantes que íbamos a necesitar con contraseñas de longitud fija, resultando 46 usuarios requeridos. Dado que se contabilizaron 66 voluntarios para las pruebas, se determinó que 70% de los participantes iban a ser escogidos de manera aleatoria para escoger una contraseña de longitud fija y los restantes podrían elegir una contraseña libre.

La cuarta etapa de recopilación del tecleo del usuario consistió en varias sesiones donde se empleó la herramienta para capturar los registros de los voluntarios inscritos. Cada semana, los usuarios tenían que hacer al menos dos sesiones, las cuales estaban espaciadas de manera inter diaria para evitar el "sobre-acostumbramiento" en la forma de teclear, pero que permitiera registrar la variabilidad del usuario. Cada sesión constaba de un flujo de tres tareas, el cual el usuario tenía que realizarlo tres veces, como se puede apreciar en la Figura 1. En la primera tarea, el usuario como usuario legítimo, tenía que iniciar sesión en el sistema con su nombre de usuario y contraseña (propia o impuesta dependiendo del grupo que le tocó). Posteriormente, se pasaba a una segunda y tercera pantalla, donde la tarea consistía en escribir tres veces, tanto el nombre de usuario y contraseña de otro usuario seleccionado de manera aleatoria. Una vez terminado este flujo, el usuario tenía que cerrar sesión, para repetir el flujo dos veces más. Se tiene que mencionar que cada registro, para que fuera considerado válido, el usuario tenía que escribir correctamente las credenciales a la primera y no se admitían correcciones, como el usar la tecla "backspace" para escribir nuevamente. Si es que el usuario se equivocaba o presionaba la tecla "backspace", la misma herramienta automáticamente borraba en pantalla lo ingresado y se le indicaba con una alerta de que tenía que realizar el paso nuevamente. Cada tecla que el usuario ingresaba era capturada por el keyloguer, incluyendo las teclas especiales tales como el "shift", el "bloq mayus", entre otras. Para la validación de las credenciales, la herramienta comparaba la palabra ingresada versus las credenciales que se registraron en el sistema distinguiendo entre mayúsculas y minúsculas. La dinámica de la sesión finalizaba cuando la herramienta contabilizaba la cantidad de muestras correctas requeridas en base a la metodología planteada. Con respecto a las muestras "inválidas", también se almacenaron como parte del dataset, puesto que el error podría considerarse como un feature relacionado a la identidad del usuario y podría ser usado en análisis futuros. El periodo de recopilación de datos tuvo una duración de seis semanas, no obstante, la herramienta pudo seguir capturando muestras después de este periodo.

Una vez terminado el periodo de recopilación de las muestras, se procedió a generar los distintos tipos de features que formaron parte de los vectores de tiempo usados para el entrenamiento de los modelos, los cuales encontramos: el "down-down key feature", el "down-up key feature", el "up-down key feature", el "up-up key feature" y el "total time". Los primeros cuatro tipos de features, consisten en la latencia entre las distintas teclas presionadas, ya sea en el momento en que se oprimen ("down") como cuando también se liberan ("up"). El "total time" consiste en todo el tiempo en que se demora el usuario en escribir la palabra. Los modelos de clasificación no pueden trabajar directamente con los datos en crudo generados por la herramienta dado que este solo captura el timestamp cuando la tecla es presionada y liberada. Esto se puede apreciar en la Figura 19, donde el usuario "5f77912b203*" escogió un username de 8 caracteres y la primera letra presionada es la tecla "T" con un timestamp asociado de 1601671460351. Para que pueda ser usable este dataset, a la data cruda se le aplica un procesamiento para generar los features mencionados anteriormente. En la Figura 2 se muestra a mayor detalle cómo generamos estos vectores de tiempo, donde para cada tecla, capturamos el evento de down (t_d) y up (t_u). Las latencias o diferencias entre los distintos eventos permiten construir los vectores de cada usuario que luego serán usados en los modelos. Por ejemplo, tomando como referencia Figura 2, el feature down-down lo obtenemos restando el tiempo en que la tecla 1 es presionada (t_{d1}) menos el tiempo capturado de la tecla 0 (t_{d0}). Esta misma lógica de diferencias se replica para los otros features de latencia. En un estudio previamente hecho por Lo, et al. (2020), se realizó un benchmark de cuatro modelos de reconocimiento de patrones de tecleo utilizando el dataset público de GREYC. Dado que este trabajo es la continuación de la investigación anterior, en la última etapa, se midió el performance de los modelos, ya previamente trabajados usando el dataset propio generado.

Figura 2

Procedimiento para generar los features de tiempo basados en latencia



5. EXPERIMENTACIÓN

La herramienta implementada consiste en una aplicación web con cuatro pantallas y un servidor Nodejs en la nube encargado de recopilar todas las muestras enviadas por los usuarios desde el navegador. Para almacenar los datos, se empleó una base de datos no relacional desplegado en los servicios de MongoDB Atlas. Se optó por una base de datos no-sql para facilitar el almacenamiento y post-procesamiento de la estructura de datos variable de los usuarios.

La primera pantalla, como se aprecia en la Figura 13, muestra el formulario de registro de un nuevo usuario en el sistema. Esta solicita los datos personales que han sido mencionados en la metodología. En esta etapa, el usuario también tiene que definir el nombre de usuario y contraseña que utilizará a lo largo de la experimentación. Como se indicó en la metodología, existirán algunos usuarios seleccionados de manera aleatoria en una proporción de 30% que tendrán que escribir una contraseña con longitud fija, por lo que se aplicó una distribución al azar que permita distribuir los usuarios en estos dos grupos. Se tiene que mencionar adicionalmente que se implementó una especie de token digital, mediante una cookie, que se almacena en el navegador local del usuario al momento de realizarse correctamente la creación de la cuenta en el sistema. Este será usado para controlar la autenticidad de las muestras del usuario cada vez que se envíen a la base de datos.

La segunda pantalla (ver Figura 14) muestra la primera tarea que realizarán los usuarios. Estos deben de ingresar las credenciales utilizadas al momento del registro. Cada vez que el usuario inicie sesión con sus credenciales originales, estos serán enviados al servidor junto con el token único del usuario. Si los datos proporcionados por el usuario son correctos, se considerará como una muestra válida de un usuario "genuino"; caso contrario, se usará el

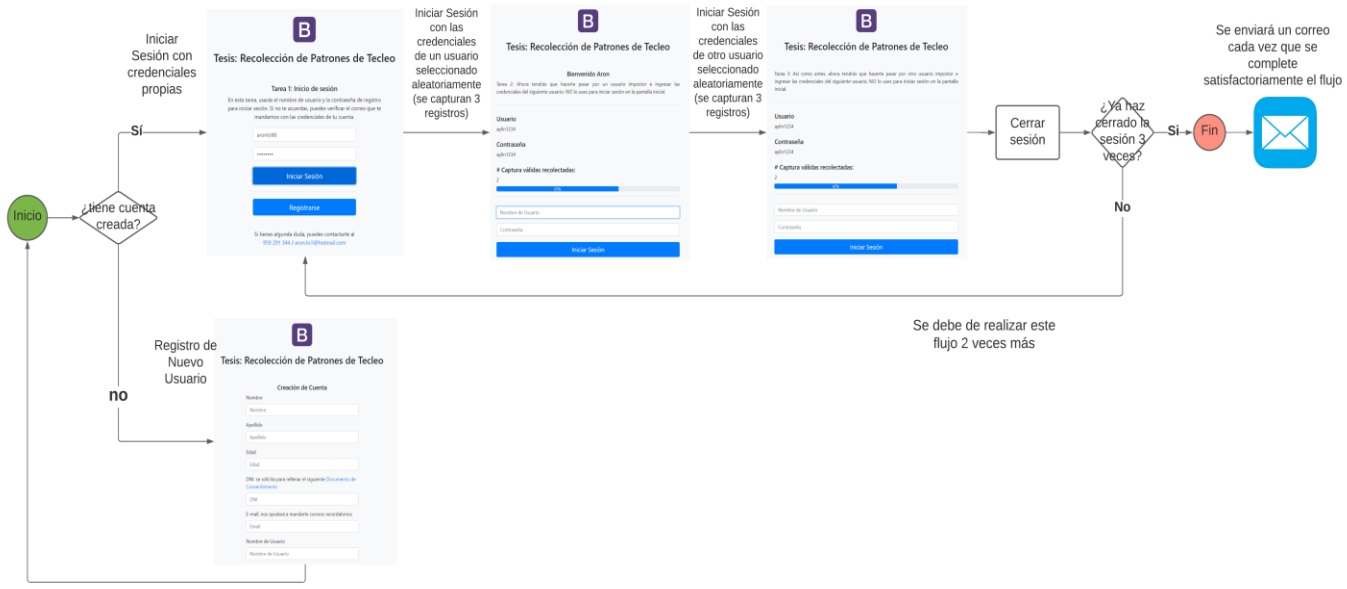
token para identificar al usuario y se considerará como un intento fallido. En ambos casos, las muestras siempre son almacenadas, pues se consideran valiosas para la investigación.

La segunda y tercera tarea que debe de hacer cada usuario se dan en la tercera (Figura 15) y cuarta pantalla (Figura 16) respectivamente. En ambas, el usuario debe de escribir tres veces las credenciales de otro usuario, el cual es proporcionada de forma aleatoria por el sistema. Un punto a considerar es que en las tres veces, el usuario debe de ingresar correctamente las credenciales mostradas al primer intento, es decir que no se admiten correcciones por parte del usuario. En este punto, por ejemplo, si el usuario decide corregir la palabra presionando la tecla “backspace”, toda la palabra será borrada por la interfase y se tendrá que escribir nuevamente. Otro punto a mencionar es que se utilizó un algoritmo de aleatoriedad por pesos usando la función “weighted” de la librería de Javascript “Chance” para priorizar aquellos usuarios que no tengan muchas muestras hechas por un usuario de tipo impostor, garantizando que estos tendrán más probabilidad de aparecer en estas pantallas. De esta forma, se garantiza que cada usuario del dataset tenga suficientes registros de este tipo. Cuando el usuario finalice la tercera y última tarea, este deberá de cerrar sesión para terminar el flujo de trabajo. Como se indicó en la metodología, en cada sesión semanal, el usuario tendrá que realizar este flujo tres veces, teniendo un total de nueve muestras diarias por sesión.

El sistema se desplegó para que se registren los usuarios y comenzarán a enviar sus muestras en una fase piloto, donde se les explicó previamente lo que consiste el experimento en una capacitación. Los usuarios han estado siguiendo el flujo mostrado en la Figura 3, donde se consiguieron muestras de varios usuarios.

Figura 3

Flujo de trabajo realizados por el usuario durante una sesión



Se estableció una estructura de datos variable usando listas dinámicas para almacenar los distintos datos, especialmente los registros de tecleo de los usuarios con contraseñas variables (marcas de tiempo y tecla presionada). Las marcas de tiempo están almacenadas en microsegundos, donde se utilizó una tecnología web mencionada en la documentacion de The Mozilla Corporation (2020) para generar marcas de tiempo de alta precisión (DOMHighResTimeStamp) con una precisión de captura de hasta cinco milésimas de segundo (5 microsegundos). En la Figura 18 se puede apreciar un ejemplo de los datos almacenados de un usuario en el dataset, donde se tienen campos importantes como el nombre del usuario, la contraseña, edad, la mano dominante que presenta, entre otros. Por otro lado, en la Figura 19 se puede apreciar un registro de la forma de teclear de un usuario. En ellas se pueden ver datos variados como el nombre de usuario y contraseña escrito, así como a quien le pertenece la muestra y a quien trató de simular.

El sistema estuvo desplegado por un periodo de seis semanas hasta el 17 de noviembre del 2020 capturando un total de 10994 de muestras de tecleos de los usuarios participantes y logrando recolectar en promedio 176 muestras por usuario. El dataset se usó para entrenar cuatro modelos siguiendo la metodología que se planteó en el trabajo de Lo et al. (2020). El trabajo previo hace uso del dataset público de GREYC (2009), el cual cuenta con una estructura de datos denominada vector de tiempo que se usa para entrenar los modelos de dinámica de tecleo. Para hacer uso del dataset que se propone en el trabajo presente, se generó un script para convertir la data cruda en estos vectores de

tiempo. Este script consiste en hacer diferencias entre las distintas marcas de tiempo de los eventos de presión y liberación de cada tecla presionada en las sesiones de recopilación para finalmente generar los siguientes features que serán usados para entrenar y testear los modelos: “down-down key feature”, el “down-up key feature”, el “up-down key feature”, el “up-up key feature”. En el trabajo de Lo et al. (2020) se usan cuatro tipos de modelos para detectar la identidad del usuario por medio de la forma de teclear, los cuales se dividen en dos categorías. El primer enfoque hace uso de las distancias como métricas de similitud, donde se emplean la distancia Euclidiana y la de Manhattan. El otro enfoque corresponde al uso de modelos de aprendizaje de máquina, donde se usaron los algoritmos de SVM y Random Forest para identificar a los usuarios.

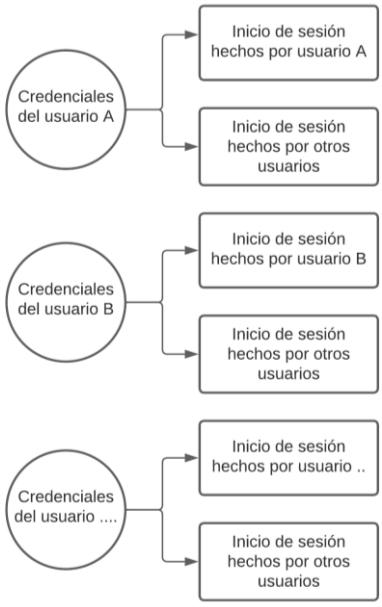
Como se mencionó en la metodología para la creación del dataset, se planteó la captura del tecleo tanto del nombre de usuario como de la contraseña en la herramienta web. Dentro del estado del arte, se menciona que las contraseñas son fijas y son impuestas hacia los usuarios. A fin de darle otro enfoque, se permitió que ciertos usuarios puedan elegir sus contraseñas manteniendo el esquema de tener una logitud fija. Así mismo, la captura de credenciales de longitud variable es algo que no suele encontrarse en la literatura, por lo que se seleccionaron de manera aleatoria usuarios para que formen parte de este grupo. De acuerdo con lo mencionado previamente en la metodología, el 70% de la población de usuarios conforman el primer grupo, donde se puso la restricción para que usen una contraseña de longitud fija al momento de registrarse en la plataforma; en cuanto al resto, se les permitió ingresar credenciales de longitud variable.

Para la generación de los modelos, siguiendo el mismo esquema que se planteó en el trabajo de Lo et al. (2020), donde se empleó el dataset de GREYC, se decidió limitar el alcance de los modelos, al entrenarlos y testearlos usando solamente las contraseñas pertenecientes a los usuarios que tienen la restricción de contraseña fija. Las muestras de los usuarios con credenciales de longitud variable, si bien no fueron usadas, forman parte del dataset generado para que pueda ser empleado en cualquier trabajo de investigación futuro.

En el dataset, con respecto a las credenciales de un usuario, como se puede ver en Figura 4, este puede dividirse en dos tipos de muestras: aquellas realizadas por el propio usuario, que pasarían a convertirse en las muestras del usuario genuino y las consideradas como impostores, cuando son digitalizadas por otros usuarios. De esta manera convertimos la clasificación en una de tipo binaria, donde independiente del modelo que se use, ya sea los basados en distancias o de aprendizaje de máquina, estos tienen que finalmente clasificar si las muestras entregadas pertenecen a un usuario genuino o impostor.

Figura 4

Estructura de los datos de las credenciales por cada usuario



Para la implementación de los distintos modelos, se usaron los mismos pasos y/o procedimientos mencionados en el artículo de Lo et al. (2020), los cuales también fueron mencionados en la sección de metodología. Los modelos basados en distancias no requirieron de ninguna configuración adicional. No obstante, los basados en aprendizaje de máquina necesitaron de un refinamiento de los hiperparámetros, dado que fueron entrenados en una investigación anterior con el dataset de GREYC. Con respecto al modelo de SVM, por medio de un “grid search”, se obtuvo los siguientes mejores parámetros: C = 100.0, gamma = 1.0, kernel = 'linear'. Para el modelo de Random Forest

se obtuvieron los siguiente: max_depth = 10, max_features = 'auto', min_samples_leaf = 1, min_samples_split = 5, n_estimators = 1300. El código fuente de las implementaciones están puestas en la sección de anexos junto con sus respectivos resultados.

6. RESULTADOS

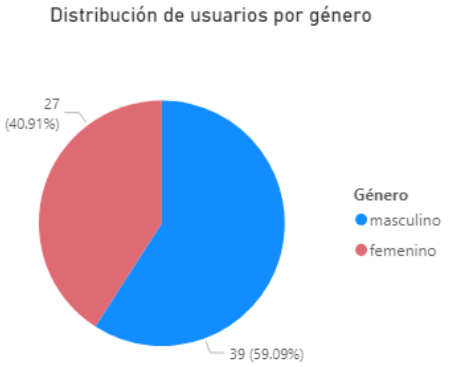
6.1 Descripción general del conjunto de datos

En esta sección, se presentará un resumen de los datos recolectados en el rango de las fechas del 2 de octubre hasta el 17 de noviembre del 2020. En lo referente a los usuarios, se contó con la participación de un total de 66 personas durante el periodo mencionado. Estos fueron reclutados por una invitación abierta enviada principalmente a la comunidad universitaria. A las personas que decidieron participar, se les contactó, se les explicó un poco acerca del trabajo de investigación, así como la dinámica del experimento, y se les brindó los accesos a la herramienta web para que puedan registrarse y comenzar a registrar sus muestras de tecleo. A continuación, se detallarán las principales características de la población participante:

Como se puede ver en la Figura 5, se puede apreciar una mayoría masculina de un 59% con respecto al total de usuarios.

Figura 5

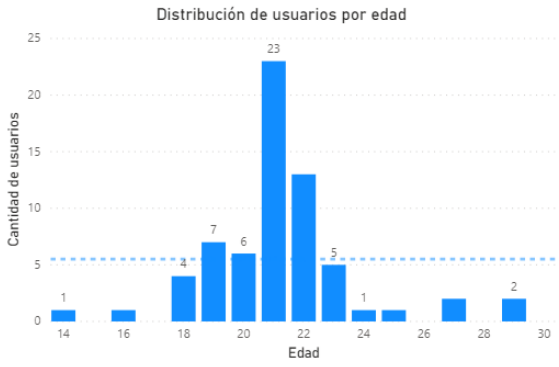
Distribución de usuarios por género



Entre todos los participantes, se puede ver en Figura 6 que las edades se concentran en personas de entre 19 a 23 años, lo cual es una población que se ubica entre la generación de adultos jóvenes.

Figura 6

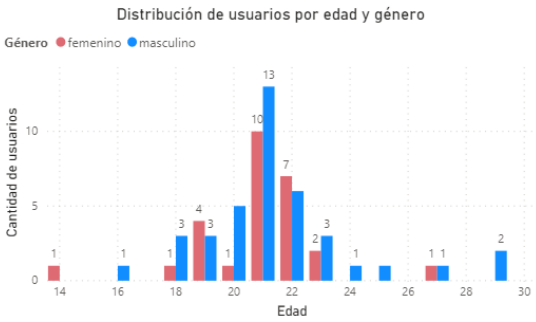
Distribución de usuarios por edad



En la Figura 7 se observa de una manera más resumida el universo de la población de participantes, pero segmentado tanto por edad y género. En todas las edades, se puede ver que existe una ligera mayor cantidad de usuarios masculinos que femeninos, especialmente en el grupo de usuarios que tienen 21 años.

Figura 7

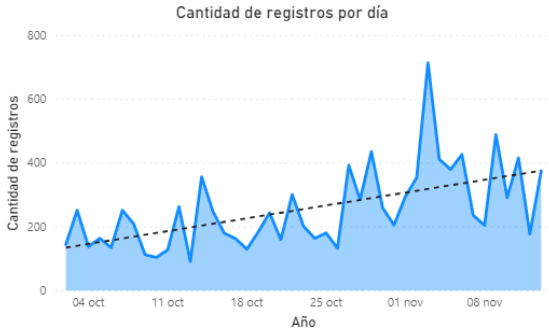
Distribución de usuarios por edad y género



Actualmente el dataset presenta un total de 10994 registros. Esta cantidad agrupa tanto los registros de inicio de sesión de usuarios impostores como los genuinos. A continuación, en la Figura 8 se puede apreciar la cantidad de registros que se han ido almacenando cada día desde que comenzó el experimento. Se puede ver un aumento considerable de registros diarios gracias a una mayor participación de parte de los usuarios, así como un incremento de nuevos usuarios en las últimas semanas de octubre.

Figura 8

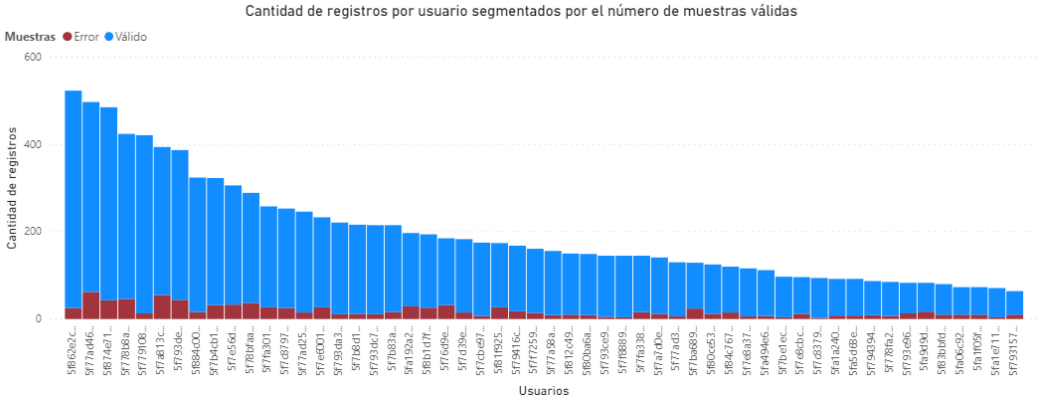
Cantidad de registros por día



Como se mencionó previamente en la metodología, los errores de inicio de sesión han sido capturados también en la herramienta para formar parte del dataset actual. Estos son considerados importantes, ya que la equivocación de teclear también puede ser considerada como parte del patrón de tecleo de uno mismo. Se realizó un análisis de la cantidad de veces que el usuario se ha equivocado en teclear las credenciales en la herramienta hasta la fecha. En la Figura 9 se aprecia la proporción de error por cada usuario, donde se observa que, en ninguno la cantidad de errores sobrepasa a la cuarta parte de la cantidad de registros totales por cada usuario.

Figura 9

Proporción de errores de inicio de sesión por cada usuario



6.2 Análisis ROC del dataset usando distintos modelos

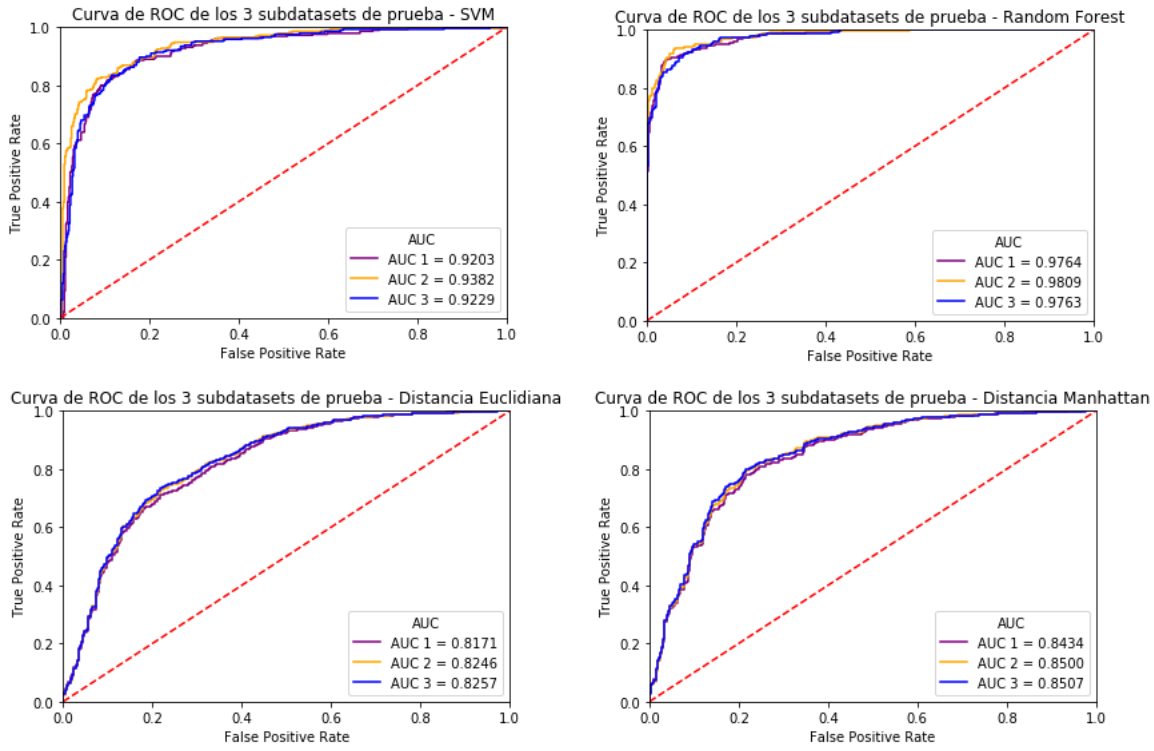
Se debe de recordar que la estructura de datos presentada en Figura 4 muestra que para cada credencial, se tiene un conjunto de muestras que fueron hechos por el mismo usuario (genuinos) y otro conjunto de muestras que fueron realizados por otros usuarios (impostores). Si observamos la Figura 1, la metodología de recolección planteada en el trabajo presenta una serie de pasos, que finalmente hará que se tenga una mayor cantidad de muestras perteneciente a usuarios impostores que genuinos en el dataset, siendo más específicos, se tendría una proporción de 3:18.

Para determinar si el dataset creado permite finalmente generar modelos que permitan distinguir entre usuarios genuinos como impostores, incluso con la cantidad de muestras desbalanceadas, pues se tiene una mayor cantidad de registros de impostores por cada usuario, se realizó un análisis haciendo uso de la curva ROC. Para evaluar de manera imparcial, sin que haya una desproporción de las muestras entre los genuinos e impostores, se aplicó el enfoque que se muestra en la sección de anexos Figura 20 y Figura 21. Este consiste en dividir el conjunto de muestras de impostores en 3 subsets de manera aleatoria, mientras que el conjunto de datos de los genuinos se usa en su totalidad. Se entrenan y se evalúan tres modelos en paralelo, donde en cada uno de los modelos se usa el mismo subset de muestras genuinas, pero tomando diferentes subsets de impostores. En cuanto al conjunto de datos de entrenamiento como evaluación, se hace uso de una proporción de 80/20. El resultado final sería la obtención de tres curvas ROC, donde se tendría que esperar que fueran similares, independientemente del subset de impostores utilizado.

Las pruebas se hicieron usando cuatro modelos distintos. Como se observa en la Figura 10, las curvas son similares independiente del subset de impostores usados. Este comportamiento se replica en los cuatro modelos presentados. Con esto, se puede mostrar que es factible modelar el dataset tanto para muestras de tipo genuinas como de impostores, permitiendo la creación de cualquier modelo que reconozca estos dos tipos de clases.

Figura 10

Análisis ROC del dataset usando distintos modelos



6.3 Métricas de desempeño

Una vez comprobado la factibilidad de modelar las muestras en las clases de usuarios genuinos como de impostores, se llevó a cabo el entrenamiento y evaluación de distintos modelos de dinámica de tecleo, los cuales ya fueron previamente implementados haciendo uso de un dataset distinto. Estos están documentados en el trabajo de Lo (2020).

En la Tabla 1 se muestran las distintas métricas obtenidas de cada modelo para determinar sus respectivos desenvolvimientos, entre los cuales tenemos: el accuracy, el recall, el f1-score y el AUC. Se puede observar que los

modelos tanto el de Random Forest como SVM, que se basan en modelos de aprendizaje de máquina, presentan un mejor desempeño con respecto a cualquier modelo basado en distancias.

Tabla 1

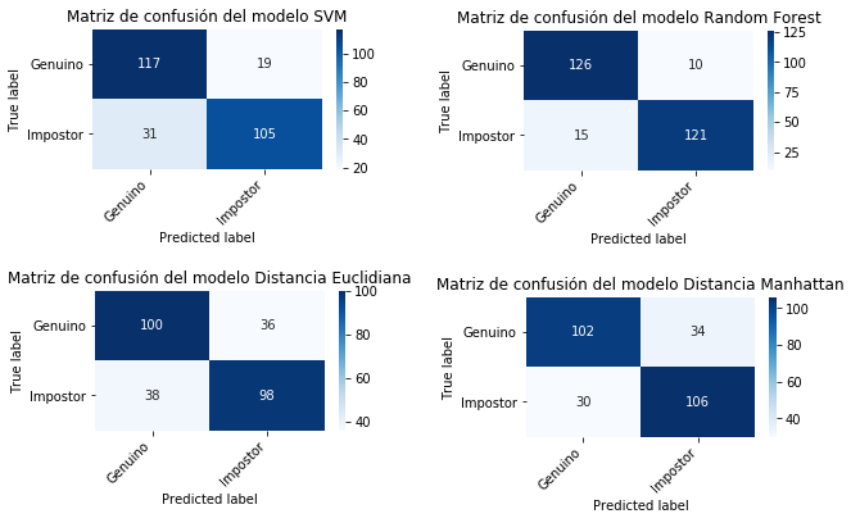
Métricas de rendimiento entre los distintos modelos

Modelo	Accuracy	Recall	F1	AUC
Random Forest	0.9081	0.9265	0.9097	0.96
SVM	0.8162	0.8603	0.8239	0.87
Manhattan	0.7647	0.7500	0.7612	0.83
Euclidiano	0.7279	0.7353	0.7299	0.80

En la Figura 11 se presentan las matrices de confusión de los modelos usados para probar el dataset creado. Se puede observar que es en los modelos de distancia, donde ocurre la mayor cantidad de errores de clasificación, tanto para registros genuinos como impostores. Los modelos que utilizan puntajes probabilísticos presentan mayor exactitud de clasificación, especialmente el Random Forest.

Figura 11

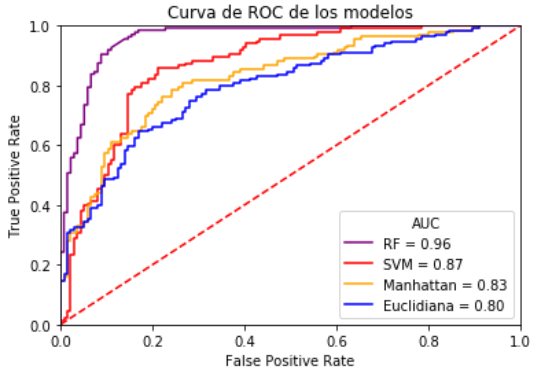
Matriz de confusión de los modelos



Si vemos las curvas de ROC en Figura 12, los modelos tanto de Random Forest como SVM presentan las curvas que están más cercanas a la esquina ideal superior izquierda, a los cuales se les asocia un AUC de 0.96 y 0.87 respectivamente.

Figura 12

Curva ROC de los modelos



7. DISCUSIÓN

El objetivo de esta investigación era desarrollar una metodología que permita la creación de un nuevo dataset de dinámica de tecleo empleando la herramienta web que se implementó para esta investigación. A partir de este dataset recolectado, se buscaría entrenar modelos de clasificación que ya previamente se han implementado en la literatura y así poder hacer una medición del performance de estos.

Mediante un análisis usando la curva ROC, se pudo comprobar que se puede hacer un modelado entre las clases genuinas como impostores usando las muestras de los distintos usuarios del dataset que se creó usando la herramienta web.

En la investigación anterior de Lo et al. (2020) se realizó un trabajo usando distintos modelos de la literatura, los cuales fueron entrenados y evaluados haciendo uso del dataset público de GREYC. Este dataset a diferencia del propuesto en el presente trabajo, cuenta con 133 usuarios de los cuales, donde todos teclearon la palabra fija "greyc laboratory", el cual es una contraseña con longitud fija. Además, por cada usuario, se tiene entre 51 a 100 muestras de tecleo. El dataset de este trabajo, el cual aún se encuentra en etapa de recolección de datos, cuenta con la participación de 66 usuarios, los cuales presentan una cantidad de muestras más variadas por usuarios debido a que muchos comenzaron en periodos distintos. Este rango de muestras puede oscilar entre 3 a 78 muestras por usuario. Para el entrenamiento de los mismos modelos por cada usuario, se usó del dataset total, aquellas contraseñas que presentan también longitud fija. La diferencia de este último es que pese a ser de longitud fija, las credenciales de los usuarios son distintas.

A continuación, se presenta en la Tabla 2 una comparación entre las distintas medidas de precisión obtenidas para los mismos modelos, pero usando dos datasets distintos. Se puede observar que, en todos los casos se tienen un accuracy por encima del 72%; no obstante, si lo comparamos entre datasets, los modelos que fueron entrenados y evaluados haciendo uso del dataset GREYC presentaron mayores niveles de precisión. Otro punto importante es que a pesar de que los modelos fueron entrenados haciendo uso de muestras de tecleo de otros usuarios, estos presentan el mismo orden de precisión entre los modelos.

Tabla 2

Accuracy entre los modelos usando dos datasets

Modelo	Accuracy - GREYC	Accuracy - Propio
Random Forest	0.9565	0.9081
SVM	0.9186	0.8162
Manhattan	0.7861	0.7647
Euclidiano	0.7440	0.7279

De acuerdo con los resultados, existe una diferencia significativa donde se demuestra que los modelos que emplean un enfoque basado en aprendizaje de máquina como el de Máquinas de Soporte Vectorial o Random Forest, resultan ser capaces de verificar de mejor manera las identidades de los usuarios independientemente del dataset usado. Esto permite corroborar las conclusiones obtenidas en el trabajo anterior.

Existen varias razones del porqué los modelos tendrían menores resultados usando el dataset que se ha creado, pese a que siguen la misma metodología y/o procedimientos. A diferencia del dataset de GREYC, la cantidad de usuarios resulta mucho mayor que el que se ha creado en este experimento, además de que la cantidad de registros por usuario es significativamente mayor. Otro punto a adicionar es que las pruebas de recolección, a diferencia del de GREYC, se dieron en un entorno web no controlado, lo cual hace que los datos de tecleo de los usuarios sean más variados a lo largo del tiempo. También, el tamaño y variabilidad de la contraseña durante el experimento puede ser un factor que haya podido influenciar en los resultados de los modelos. Mientras que el dataset de GREYC usaba una misma palabra para todos los usuarios, el dataset propio presenta un conjunto de contraseñas distintas para cada usuario que pese a tener longitud fija, la palabra tecleada es diferente para cada usuario. Esto último puede aumentar el factor de la complejidad del mismo dataset.

Por el momento, los resultados usando nuestro dataset parecen prometedores; no obstante, ya que aún se están recolectando muestras de usuarios, es muy difícil sacar conclusiones totalmente cerradas acerca del dataset que se está elaborando.

8. CONCLUSIONES

La dinámica de tecleo implica analizar el patrón de tecleo de una persona para verificar si la identidad de la persona corresponde a la quien dice ser esta misma. Existen distintos datasets públicos que se pueden usar para realizar este tipo de biometría aplicando distintas técnicas de dinámica de tecleo obteniendo resultados notables en varios trabajos presentados en la literatura.

Uno de los puntos en comunes de estos datasets utilizados radica en la simplicidad de las credenciales de los usuarios, pues simplemente se les solicita escribir la misma palabra durante varias sesiones a la semana, además que se puede añadir que las recolecciones se hacen en laboratorios controlados.

De lo planteado en la investigación, se logró la creación de un dataset de dinámica de tecleo por medio de una herramienta web desplegada por varias semanas. Esta, a diferencia de otras, contiene muestras con contraseñas únicas por cada usuario. Al estar desplegada de forma web, permite que la captura de las muestras sea realizada en las mismas computadoras personales de los usuarios, lo que permite tener datos en un contexto más realista.

Finalmente, el uso de este dataset junto con antiguos modelos de dinámica de tecleo resultan prometedores, pues arrojan resultados casi parecido a los trabajos previos que hicieron uso de otros datasets públicos como el de GREYC, los cuales presentaron una precisión por encima del 72%.

9. CONTRIBUCIONES

En cuanto a la contribución del trabajo presente, el alumno Aron Lo se encargó de la conceptualización del proyecto, así como de llevar a cabo la experimentación. Por parte de los asesores, el profesor Juan Gutierrez-Cardenas participó activamente en todas las etapas del proyecto. El profesor Victor H. Ayma contribuyó en la definición y en el afinamiento de la metodología planteada.

10. TRABAJOS FUTUROS

Es importante indicar que este dataset va a ir creciendo y mejorando en cuanto a la calidad, pues los resultados presentados en este documento fueron hechos en base a una captura de la base de datos hasta cierto momento. Dado que la herramienta web seguirá desplegada, habrá una mayor cantidad de datos de los usuarios que, por ende, mejorarán los modelos que se usen en un futuro.

REFERENCIAS

Thomas, J. E. (Abril de 2018). Individual Cyber Security: Empowering Employees to Resist Spear Phishing to Prevent Identity Theft and Ransomware Attacks. *International Journal of Business and Management, 13*(6), 1.

Lisa, R. (9 de Diciembre de 2014). *Cyber attack could cost Sony studio as much as $100 million*. Recuperado el 2020, de Reuters.

Darabseh, A., & Namin, A. S. (2016). On Accuracy of Classification-Based Keystroke Dynamics for Continuous User Authentication. *Proceedings - 2015 International Conference on Cyberworlds, CW 2015*, 321-324.

Dee, T., Richardson, I., & Tyagi, A. (2019). Continuous transparent mobile device touchscreen soft keyboard biometric authentication. *Proceedings - 32nd International Conference on VLSI Design, VLSID 2019 - Held concurrently with 18th International Conference on Embedded Systems, ES 2019*, 539-540.

Monrose, F., & Rubin, A. (Febrero de 2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems, 16*(4), 351-359.

Baynath, P., Sunjiv Soyjaudah, K., & Heenaye-Mamode Khan, M. (Julio de 2018). Machine Learning Algorithm on Keystroke Dynamics Pattern. *Proceedings - 2018 IEEE Conference on Systems, Process and Control, ICSPC 2018*, 11-16.

Çeker, H., & Upadhyaya, S. (Diciembre de 2015). Enhanced recognition of keystroke dynamics using Gaussian mixture models. *Proceedings - IEEE Military Communications Conference MILCOM*, 1305-1310.

Darabseh, A., & Namin, A. (2016). Effective user authentications using keystroke dynamics based on feature selections. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, 307-312.

Zhong, Y., Deng, Y., & Jain, A. (Junio de 2012). *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 117--123.

Zhong, Y., & Deng, Y. (Febrero de 2015). A Survey on Keystroke Dynamics Biometrics: Approaches, Advances, and Evaluations. *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics*, 1-22.

Chen, C. L., Weng, K., & Chee, P. (2007). Keystroke patterns classification using the ARTMAP-FD neural network. *Proceedings - 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2007*, 61-64.

Kang, P., Hwang, S., & Cho, S. (2007). Continual retraining of keystroke dynamics based authenticator. En *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (Vol. 4642 LNCS, págs. 1203-1211). Berlin: Springer.

Giot, R., El-Abed, M., & Rosenberger, C. (2009). IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems, BTAS 2009. *GREYC keystroke: A benchmark for keystroke dynamics biometric systems* (págs. 1-6). IEEE.

Bleha, S., Slivinsky, C., & Hussien, B. (1990). Computer-Access Security Systems Using Keystroke Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(12), 1217-1222.

L. C., A., Sucupira, L. H., Lizarraga, M., Ling, L., & Yabu-Uti, J. (Febrero de 2005). User authentication through typing biometrics features. *IEEE Transactions on Signal Processing, 53*, 851-855.

Hosseinzadeh, D., & Krishnan, S. (2008). Gaussian mixture modeling of keystroke patterns for biometric applications. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 38*(6), 816-826.

Killourhy, K., & Maxion, R. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. *Proceedings of the International Conference on Dependable Systems and Networks*, 125-134.

Abramson, M., & Aha, D. (2013). User authentication from web browsing behavior. *FLAIRS 2013 - Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference*, 268-273.

Lo, A., Ayma, V. H., & Gutierrez-Cardenas, J. (2020). A Comparison of Authentication Methods via Keystroke Dynamics. *2020 IEEE Engineering International Research Conference (EIRCON)* (págs. 1-4). Lima: IEEE.

The Mozilla Corporation. (09 de Octubre de 2020). *Referencia de la API Web MDN*. Obtenido de https://developer.mozilla.org/es/docs/Web/API/Performance/now

ANEXOS

Figura 13

Pantalla de creación de cuenta del usuario

Figura 14

Pantalla de de inicio de sesión como usuario legítimo (tarea 1)

Figura 15

Primera pantalla de inicio de sesión como usuario impostor (tarea 2)

Figura 16

Segunda pantalla de inicio de sesión como usuario impostor (tarea 3)

Figura 17

Tercera pantalla de inicio de sesión como usuario impostor (tarea 4)

Figura 18

Ejemplo de la estructura de un usuario registrado en el dataset

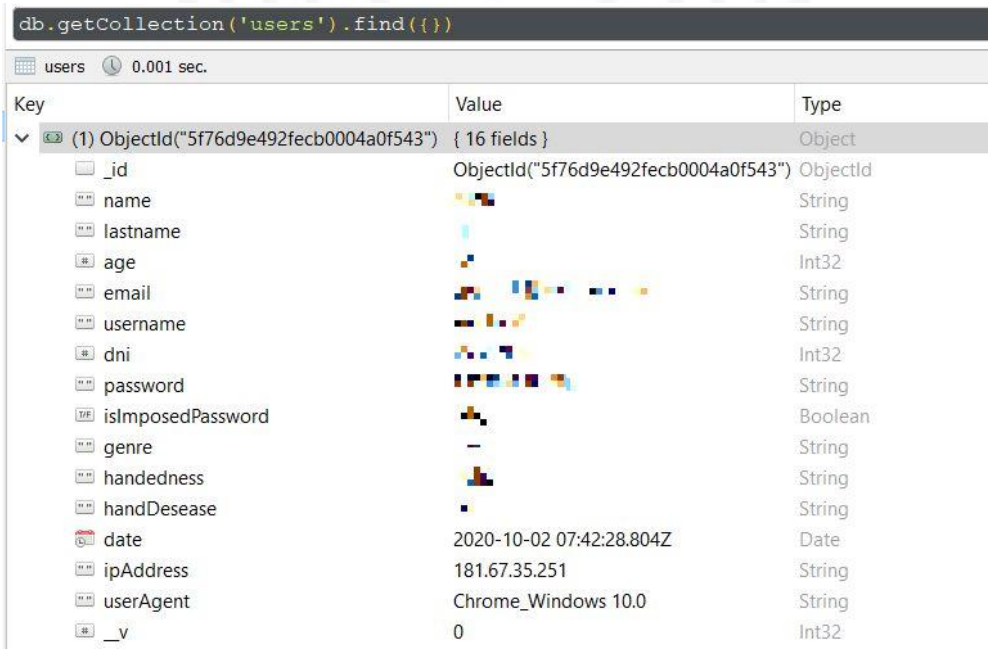


Figura 19

Ejemplo de la estructura de una muestra de tecleo registrado en el dataset

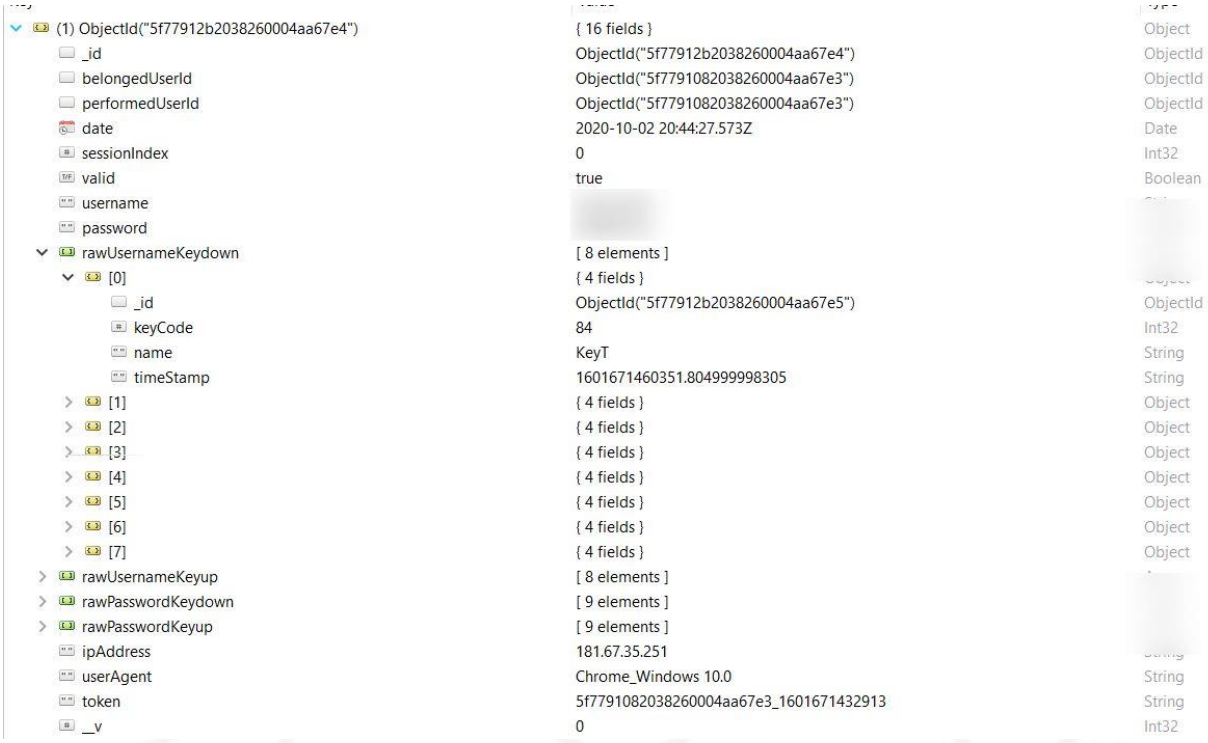


Figura 20

Análisis ROC del dataset usando modelos de distancia

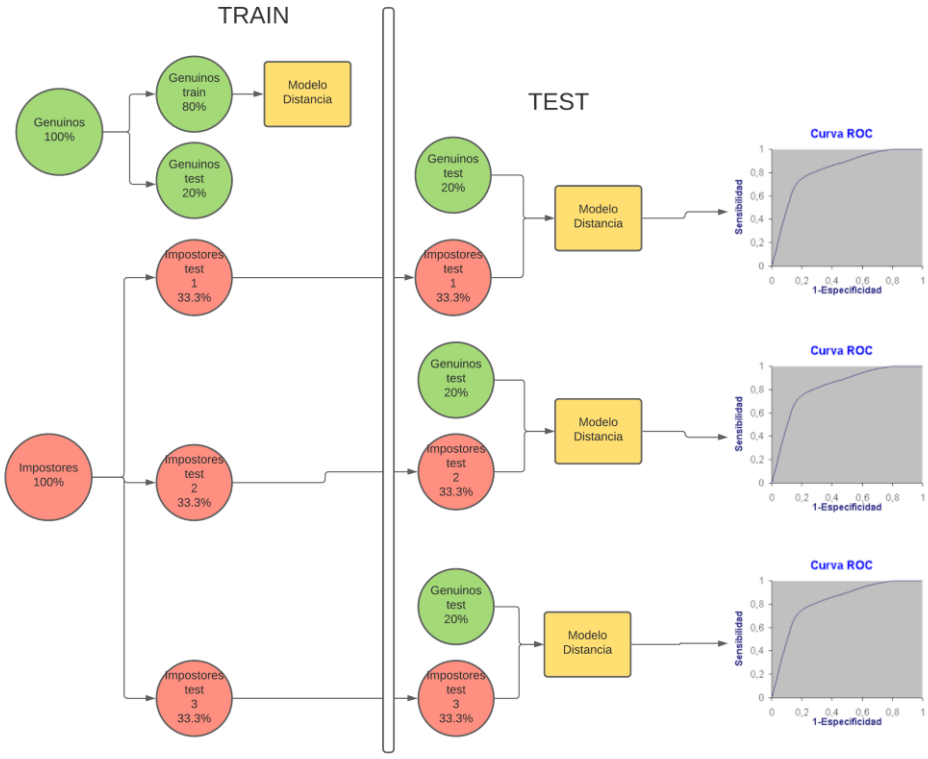
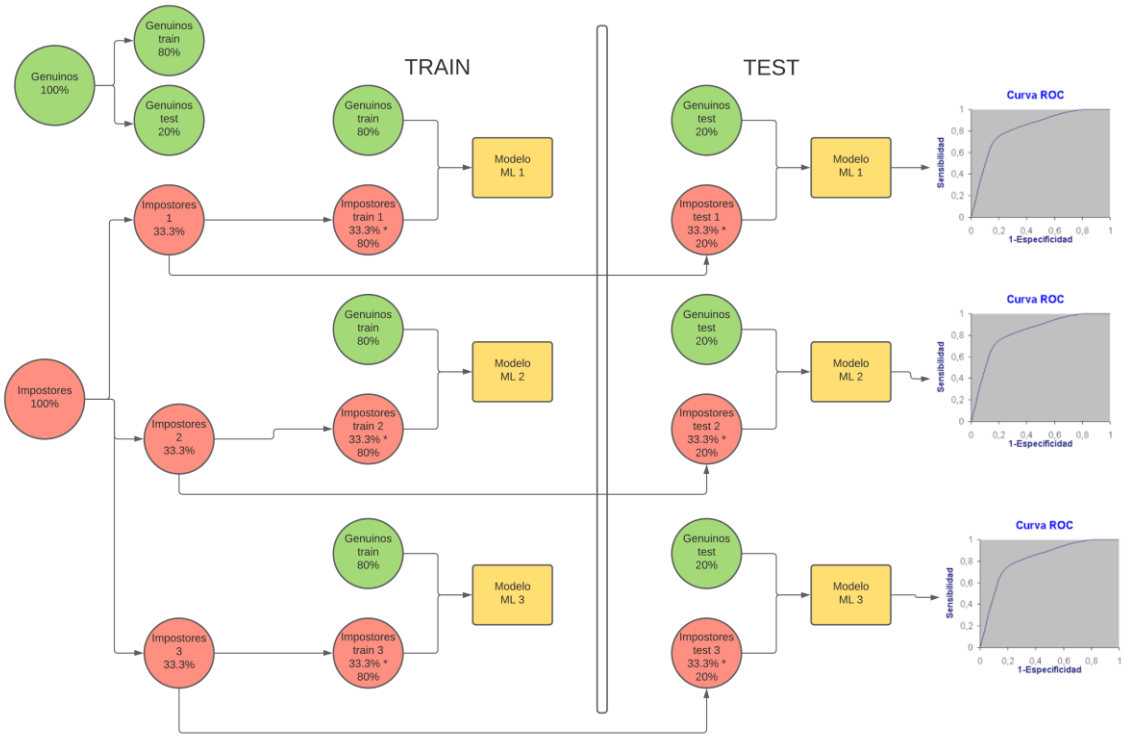


Figura 21

Análisis ROC del dataset usando modelos de aprendizaje de máquina

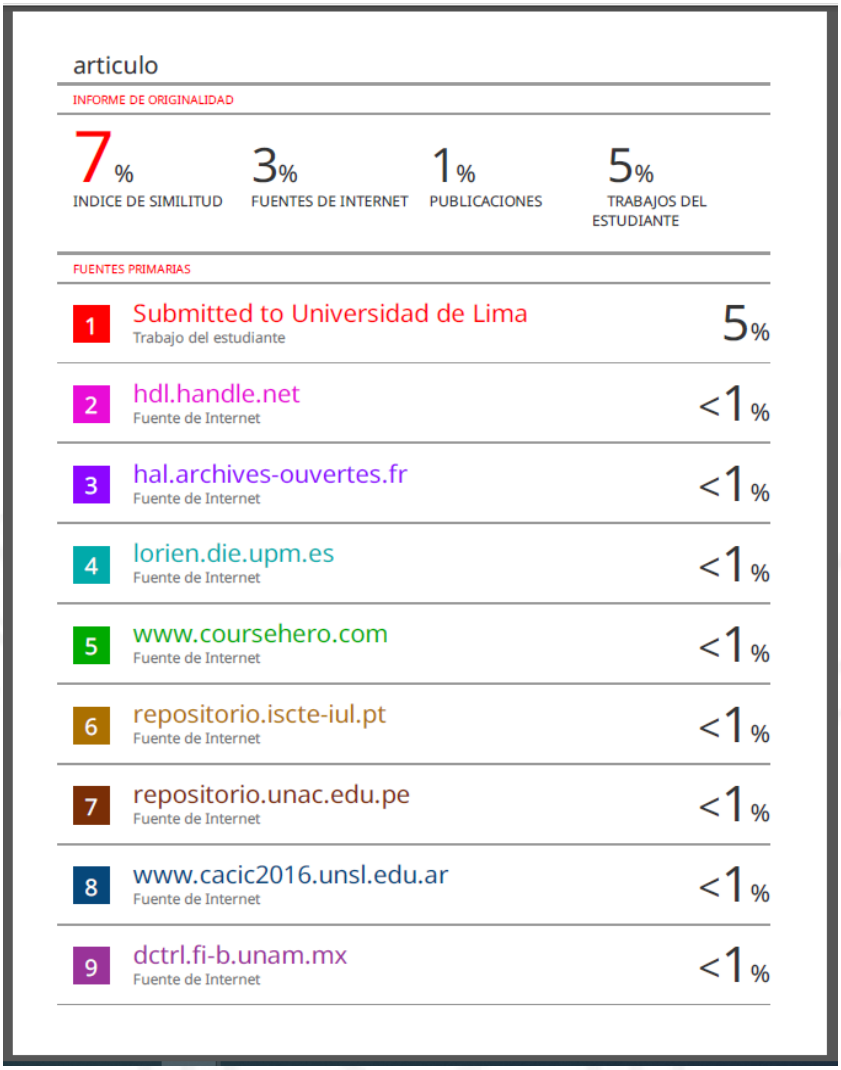


Enlace 1

Código fuente de la herramienta web implementada para la generación del dataset: https://github.com/aronlo98/tesis-keylogger

Enlace 2

Código fuente de los modelos de dinámica de tecleo: https://github.com/aronlo98/tesis



articulo

INFORME DE ORIGINALIDAD

7%	3%	1%	5%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad de Lima Trabajo del estudiante	5%
2	hdl.handle.net Fuente de Internet	<1%
3	hal.archives-ouvertes.fr Fuente de Internet	<1%
4	lorien.die.upm.es Fuente de Internet	<1%
5	www.coursehero.com Fuente de Internet	<1%
6	repositorio.iscte-iul.pt Fuente de Internet	<1%
7	repositorio.unac.edu.pe Fuente de Internet	<1%
8	www.cacic2016.unsl.edu.ar Fuente de Internet	<1%
9	dctrl.fi-b.unam.mx Fuente de Internet	<1%