VOL 3 Nº 2 Septiembre - Febrero 2022 ISSN Nº 2708-0935

Innovación y Software

Facultad de Ingeniería
Universidad La Salle, Arequipa, Perú
facin.innosoft@ulasalle.edu.pe
https://revistas.ulasalle.edu.pe/innosoft

ARK: ark:/42411/s9
PURL: 42411/s9

Vol. 3 N°. 2 2022 Septiembre - Febrero

ISSN N°: 2708-0935

DOI: 10.48168/innosoft.s9

ARK: ark:/42411/s9

PURL: 42411/s9

Depósito Legal: 2023-08884

Periodicidad: Semestral

Publicado: 30/09/2022

Editado por:

Universidad La Salle

RUC: 20456344004

Av. Alfonso Ugarte N° 517, Cercado, Arequipa

COMITÉ EDITORIAL

EDITORIAL

ARTÍCULOS ORIGINALES

La Revista Innovación y Software de la Facultad de Ingeniería, en la Universidad La Salle, se complace en presentar este segunod número de su tercer volumen que tiene como objetivo el promover investigaciones, los cambios y usos de nuevos elementos tecnológicos y su interrelación con la Ingeniería de Software y la Ciencia de la Computación.

Durante una década, hemos estado rastreando el aumento y eventual aumento de la experiencia digital, el análisis, la nube, la realidad digital, la cognición, la cadena de bloques, el negocio de TI, el riesgo y la modernización central. La actualización de este año revisa la adopción de estas macrofuerzas por parte de la empresa y explora cómo están dando forma a las tendencias tecnológicas que prevemos afectarán los negocios en los próximos 18 a 24 meses. Para hacer realidad la promesa completa de estas fuerzas, las organizaciones están explorando cómo se cruzan para crear más valor y nuevas formas de administrar la tecnología y las funciones tecnológicas. Este paso necesario se vuelve cada vez más importante a medida que las empresas se preparan para abordar las fuerzas emergentes que están en el horizonte: experiencia ambiental, inteligencia exponencial y cuántica.

En una tendencia creciente, las empresas líderes se están dando cuenta de que cada aspecto de su organización afectado por la tecnología representa una oportunidad para ganar o perder la confianza. En lugar de ver la confianza como una cuestión de cumplimiento o de relaciones públicas, la ven como un objetivo empresarial clave a perseguir. Con esto en mente, la confianza se convierte en un esfuerzo de 360 grados para garantizar que las organizaciones trabajen juntas en múltiples aspectos de la tecnología, los procesos y las personas para mantener los altos niveles de confianza que esperan sus muchas partes interesadas. Los líderes empresariales están reevaluando cómo sus productos, servicios y las decisiones que toman, en torno a la gestión de datos, la creación de ecosistemas de socios y la capacitación de los empleados, entre otros, generan confianza.

Comité Editorial

Gestión de información de lecturas del consumo de agua potable mediante una aplicación móvil

Management of drinking water consumption reading information via mobile application

Angel Geovanny Rochina Chisag
Ministerio de Educación del Ecuador.
Quito, Ecuador.
rochitheonly@gmail.com
https://orcid.org/0000-0002-1570-9624

Edwin Wilfrido Guashpa Pasto
Investigador independiente.
Ecuador.
ewguashpa@gmail.com
https://orcid.org/0000-0002-6030-3359

Jesus Antonio Coloma Garofalo
Universidad Estatal de Bolivar. San Pedro de Guaranda, Ecuador.
jcantinio24@gmail.com
https://orcid.org/0000-0003-1827-3296

ARK: ark:/42411/s9/a55
PURL: 42411/s9/a55

RECIBIDO 15/05/2022 • ACEPTADO 10/07/2022 • PUBLICADO 30/09/2022

RESUMEN

El artículo evidencia la implementación de una aplicación móvil para la gestión de información sobre el consumo del líquido vital en los hogares que conforman la junta administradora de agua potable de la parroquia Santa Fe, tal sistema informático permite mejorar los servicios en la recolección, registros, actualización de datos sobre el consumo de agua y los cobros mensuales, visualiza la ubicación geográfica de los medidores en un mapa y sincroniza la información de los usuarios en tiempo real. El estudio realizado es una investigación empírica de tipo descriptivo – aplicativo con una modalidad cuantitativa, para la gestión del proyecto y desarrollo del software se empleó la metodología ágil Scrum que permitió ejecutar las tareas de manera eficaz conjuntamente con la participación activa del cliente. Los resultados preliminares revelan que se optimizo los procesos en el transporte de la información desde la lectura de medidores que realiza el empleado de la empresa hasta su facturación (entrega de recibo al consumidor) reduciendo los tiempos de respuestas, por lo que anteriormente lo realizaban a través de fichas impresas y cuaderno de apunte para su posterior digitación e impresión. En otros proyectos similares también manifiestan que el desarrollo y utilización de las herramientas tecnológicas para el cobro de las

planillas del agua incrementa los beneficios y eficiencia tanto a los administradores y clientes de una junta administradora de una localidad cantonal o regional.

Palabras claves: Aplicación Móvil; Agua Potable; Dispositivos Móviles; Software; Gestión de información.

ABSTRACT

The article evidences the implementation of a mobile application to manage information on the consumption of vital liquids in the homes that make up the administrative board of drinking water of the Santa Fe parish. Such a computer system allows for improving collection services and records, updating water consumption and monthly charges data, viewing the geographical location of the meters on a map, and synchronizing user information in real time. The study is an empirical investigation of a descriptive type - application with a quantitative modality. For project management and software development, the agile Scrum methodology was used, which allowed the tasks to be executed effectively together with the client's active participation. The preliminary results reveal that the processes in the transport of the information were optimized from the meter reading carried out by the company employee to its billing (delivery of receipt to the consumer), reducing response times, for which previously they were carried out by through printed cards and notebook for later typing and printing. In other similar projects, they also state that the development and use of technological tools for collecting water bills increase the benefits and efficiency for administrators and clients of an administrative board of a cantonal or regional locality.

Keywords: *Drinking water; Information Management; Mobile app; Mobile devices; Software.*

INTRODUCCIÓN

Durante las últimas décadas hemos presenciado la evolución de las tecnologías de información y comunicación (TIC), la tecnología de teléfono móvil, y junto a ello, la tecnología Wireless Application Protocol (WAP), esto ha permitido agrupar una serie de protocolos y estándares para poder transmitir la información, WAP une dos elementos imprescindibles como es el internet y la comunicación inalámbrica, con la cual es posible tener acceso a la información o servicios de la web desde un terminal móvil (teléfono celular) inalámbrico [1]; la primera generación (1G) de las tecnologías de teléfono móvil y conexiones inalámbricas cumplían funciones básicas y elementales comparada con la con la actual 5G (quinta generación) que son multiprocesos y multitareas [2].

Los teléfonos móviles son los más utilizados en la actualidad por ser considerados como pequeños ordenadores que tienen grandes capacidades de almacenamiento y procesamiento gracias a la

combinación de sus componentes (cámaras, gps, sensores, pantallas, etc.), son de distintos tamaños y pesos, y ofrecen una diversidad de aplicaciones móviles (APP) y no paran de crecer en su entorno cada día con mejores servicios [3]; las APP facilitan la gestión de información a los usuarios, ejecutan tareas determinadas y concretas ayudando a optimizar el tiempo de trabajo, las primeras aplicaciones fueron básicas como agenda, calculadora, contactos, mensajería, entre otras, con el pasar de los años han optado por los videojuegos, multimedia, redes sociales, actividades financieras, administrativas, entre otras [4], tales aplicaciones informáticas son desarrollados para diferentes sistemas operativo como Android, iOS, Firefox OS, Ubuntu Touch, etc.; y una de las ventajas de utilizar un dispositivo móvil es la portabilidad, autonomía, acceso a internet e integración de funciones, siendo hoy una herramienta comunicativa de primer nivel [5].

Al mismo tiempo, también han revolucionado y cambiado los modelos de negocios, las formas de difundir la información, maneras de relacionarse entre personas, e incluso las modalidades de trabajo, provocando una migración hacia la tendencia de plataformas y medios digitales (radio/tv online, periódicos electrónicos, e-commerce, m-commerce, etc.) [6], cada día los cibernautas visitan páginas web mediante las aplicaciones móviles (APP) desde un teléfono móvil inteligente, haciendo que las APP lleguen a constituir un ecosistema propio con gran capacidad de comunicación sobre los contenidos digitales [7]; la era post-PC está dominado por el uso de un smartphone (teléfono inteligente) o Tablet que se ha convertido como una tecnología transcendental de comunicación, información y convergencia entre los individuos [8].

Las aplicaciones informáticas son ejecutadas en un teléfono inteligente, las APP no son de iguales características ni son del mismo tipo, cada uno tiene sus ventajas y desventajas, actualmente tenemos tres tipos que se conocen como: (1) Nativa que es específico para un único sistema operativo, (2) Web que se pueden ejecutar en diferentes dispositivos sin tener que crear varias aplicaciones, (3) Híbrida que es una combinación de las dos anteriores mencionadas [9], estas aplicaciones como pueden ser online/offline. Cada día se va diversificando para sectores como educación, turismo, transporte, industrias, etc.

El acceso a las nuevas tecnologías y el internet en los hogares cada día crece, y juegan un papel sustancial con el usuario, muchos con mayores influencias positivas y otros con influencias negativas, existe una brecha significativa entre las zonas rurales y urbanas por la facilidad de instalar la infraestructura [10]. Según los informes de [11], [12] destacan que en Ecuador más de 13.8 millones de usuarios de internet en promedio 92% han tenido una audiencia en las diferentes plataformas digitales y sociales vía dispositivos móviles, y se ha multiplicado la cantidad de usuarios que realizan transacciones online, del 2% al 10%, demostrando el potencial de mercado y oportunidad en la web.

Las empresas actualizan procesos con automatizaciones para ampliar la producción y bajos sus costos de operación y de a poco van dejando de utilizar los sistemas tradicionales, lo que con

lleva a la búsqueda de nuevas tecnologías emergentes más adecuados [13], por ello los dispositivos móviles son una innovación para procesos más rápidos y servicios de calidad, y de apoco las personas van aceptando el uso de la tecnología móvil para el pago de los servicios básicos [14] como energía eléctrica, agua, teléfono.

El agua es vital para la supervivencia de los seres humanos, la cual es un elemento importante para el desarrollo de la vida [15], la fisiología del agua es transcendental para todas formas de subsistencia por tal razón los países en el mundo dan un tratamiento especial al agua [16] mediante una planta potabilizadora (agua potable) antes de ser consumida y evitar enfermedades que pueden causar las bacterias, virus, sustancias, microorganismo que están en el líquido vital para ello se aplica múltiples estrategias, buenas prácticas apoyadas en la experiencia para mejorar la calidad [17].

Las Juntas Administradoras de agua potable (JAAP) son organizaciones barriales, comunitarias, parroquiales, cantonales y/o empresas sin fines de lucro encargadas de controlar las plantas potabilizadoras, para su subsistencia requieren de capital social, y para el mantenimiento de la infraestructura en muchos casos aplican prácticas de solidaridad (minga) y truques entre los usuarios, esto les permite enfrentar problemas socioeconómicas y demográficos [18]; es necesario la instalación de medidores (dispositivo o artefacto) para registrar con precisión la cantidad del líquido (agua) y evitar el mal uso (derroche) en cada uno de los hogares, además esto ayuda a tener una mayor seguridad, control y equilibrio del consumo de agua [19]; los administradores de la JAAP junto a los usuarios también tienen la tarea de proteger , recuperar y conservar las fuentes de agua y del manejo de los páramos.

La Junta de agua potable de la parroquia Santa Fe utiliza un proceso manual para la recolección de datos de consumo del agua y su facturación, son anotados en hojas pres impresos o cuadernos apuntes que en muchas ocasiones son llenados erróneamente en casos subiendo o bajando la cantidad de agua consumida, los comprobantes de la facturas y listado de los socios se encuentran desorganizada y desactualizada en los folders y existe una demora en su búsqueda por la no disponibilidad inmediata de los datos, y corren el riesgo de desaparecer por algún desastre natural [20] como lluvias, terremotos, incendios o falla humana, un operador (persona encargada de ir de casa en casa) nuevo al ingresar a laborar desconoce las rutas y su ubicación geográfica por lo que lleva mucho tiempo en encontrar al usuario lo cual ha provocado descontento e inconformidad por el servicio, los procesos en la búsqueda y visualización de datos son repetitivos cada mes por su acumulación de archivos físicos y en el programa Excel siendo una mescla en su almacenamiento, lo cual ha llevado a descuadres financieros, inexistencia general o individual de reportes, demoras en las recaudaciones o con valores económicos altos por los registros erróneos de las lecturas y por ende quejas continúas parte de los contribuyentes de la JAAP.

Con el proyecto se pretende mejorar los procesos de registro de los usuarios, tomas de lecturas, generación de planillas, su almacenamiento, evitar duplicidad en el registro y la elaboración de

informes en la JAAP haciendo el uso de las herramientas tecnológicas en este caso una aplicación móvil, con el cual se reducirá los tiempos de respuestas, sincronización en tiempo real de datos, con una mejor gestión de la información y dar un ágil y eficiente atención a los clientes e incluso se podrá utilizar sin la necesidad de conexión a internet, el sistema será desarrollado acorde a los requerimientos de los socios y aplicara el plan de pruebas (modelo V) para garantizar la calidad [21] y para facilitar una mejor experiencia del usuario en la navegación.

Los beneficiarios directos serán los administradores y los beneficiarios indirectos todos los usuarios de la junta de agua potable de la Parroquia Santa Fe perteneciente al cantón Guaranda provincia Bolívar, con la investigación se podrán elevar el nivel en la calidad del servicio a los usuarios, así, como también evitar la pérdida de información.

Materiales y métodos

Para la implementación del proyecto se acudió a la metodología de investigación específicamente con el tipo de investigación descriptiva que ayudó a describir la problemática que se encuentra atravesando la junta administradora de agua potable (JAAP) [22], además de realizar un estudio libre de cada característica registrada, de forma acentuada, por lo que fue posible de una u otra forma incorporar las estimaciones de al menos dos cualidades cuantitativos y cualitativos para su análisis, con la exploración de campo nos consintió recolectar información de las diferentes actividades con el fin de examinar sobre las fallas e inconvenientes que posee la empresa, las cuales se aplican de manera práctica y directa al equipo administrativo de la JAAP. Y se complementó con la revisión bibliográfica en libros, revistas, periódicos, blogs y artículos publicados en sitios web, para compilar la información relevante al tema de estudio y sustentar dentro del marco referencial, conceptual y en partes que se toma para justificar las teorías del trabajo desarrollado.

En la selección de los datos de la exploración para levantar información de manera inmediata se aplicó las Técnicas e Instrumentos [23] como: *la entrevista* que se ejecutó de manera oral y verbal a los directivos de la JAAP involucrados en el desarrollo del proyecto, se usó *la encuesta* que fue un cuestionario previamente diseñado para obtener los requerimientos a ser tratados en la aplicación móvil y finalmente *la observación directa* el mismo que nos hizo comprender el flujo de información, las fallas, e inconsistencias del proceso manual que lleva en la junta administradora.

El *Universo y Muestra* [24] de la junta de agua potable de la Parroquia Santa Fe, laboran 6 personas que conforman la directiva como son el presidente/a, vicepresidente/a, secretario/a, tesorero/a, dos vocales, y una persona (empleados) quien se encargan de la recolección de lecturas de los medidores para conocer el consumo mensual de cada usuario. Por tanto, en el

proyecto de investigación por ser una empresa pequeña se consideró a las 7 personas para realizar las encuestas y/o entrevistas. Para el procesamiento de información como la recopilación y tabulación de datos, diseño de tablas y gráficos con el análisis se tomó de los resultados obtenidos de las técnicas e instrumentos aplicados.

Una vez realizado el estudio y obtener los requisitos para el desarrollo del software, y manejar de una manera eficaz la gestión del proyecto en la JAAP, se asignó un rol (scrum master, Developers, producto owner) a cada miembro del equipo como lo manifiesta la Metodología Scrum [25] como se muestra en la figura 1, y ejecuto cada etapa 1. User stories, 2. Producto backlog, 3. Sprint backlog, 4. Sprint (weeks), 5. Increment, 6. Feedback como lo indica la scrum-ágil, el mismo que fue flexible en el todo el proceso de desarrollo de software, por tal razón nuestra elección.

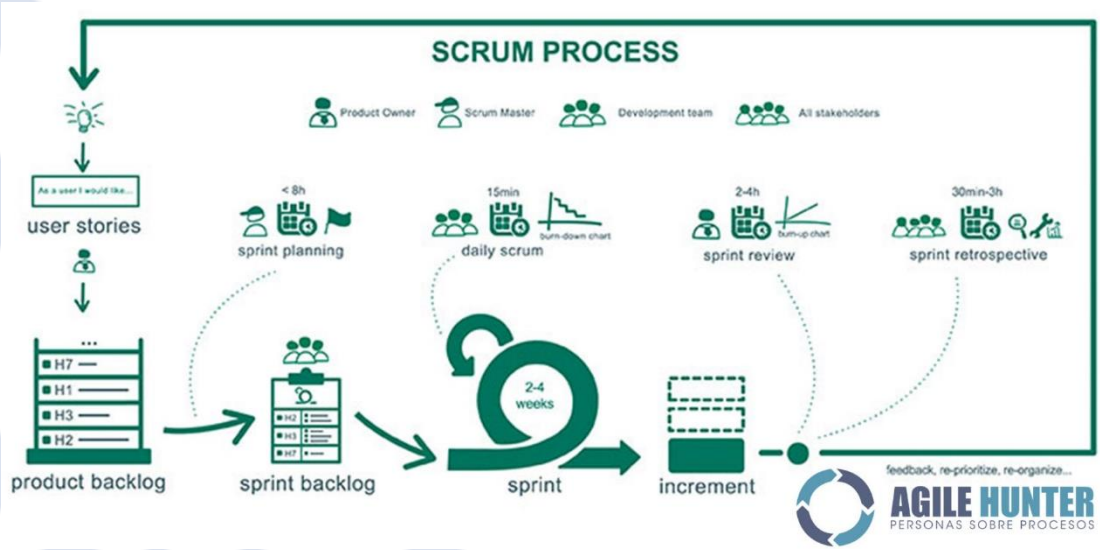

Figura 1. Procesos de la metodología scrum. Agile Hunter. (2021). Recuperado de: https://agilehunter.com/que-es-scrum/

Con la metodología Scrum se logró tener un desarrollo de aplicaciones móviles [26] muy rápidos mediante ciclos de desarrollos, que fue una gran ventaja por ser un equipo muy pequeño (de no más de diez desarrolladores) y se trabajó en un mismo espacio físico. Este método, permito conseguir productos totalmente funcionales en menos de diez semanas.

En cada fase (excepto la inicial) siempre se planteó un día de planificación y otro de entrega, y el producto se entregaba una versión prueba del software y se repitan iterativamente las sub fases para mayor calidad y asegurar el funcionamiento correctamente, con el fin de entregar una versión estable y plenamente funcional del sistema según los requisitos del cliente.

En los roles, el *Product Owner r*epresentaba al cliente y a los implicados en el proyecto de manera directa, determinando los objetivos y los propósitos para garantizar la funcionalidad del equipo

trabajo; el *Scrum Master* aseguro que el resto del equipo no tenga ninguna dificultad para cumplir con sus tareas y funciones; Developers desarrollo y entrego el producto final. En los procesos, el *Product Backlog priorizo un con*junto de requisitos (historias) de acuerdo a las necesidades; el *Sprint Backlog* en inicio a desarrollar la lista de historias; en el *Sprint* iba mostrando y verificando junto al equipo de trabajo el avance de la aplicación; en el *increment* se completaban las actividades con las sugerencias indicadas.

Resultados y discusión

De acuerdo a los objetivos planificados en el proyecto, se obtuvieron los siguientes resultados antes del desarrollo de la aplicación móvil para la administración del sistema en la Junta Administrativa de Agua Potable, en la cual participaron 6 personas (administrativo) y una persona encargada de recolectar las lecturas, observar la tabla 1 para mayor detalle.

Tabla 1. Recolección y almacenamiento de los datos de lecturas de los usuarios.

OPCIÓN	FRECUENCIA	%
Cuadernos de apuntes	1	33,3
Hojas de Excel	1	33,3
Otros (Fotos, etc.)	1	33,3

Según el estudio realizado como se muestra en la tabla 2, la persona encargada de recabar la información mensual de los medidores de agua lo realiza mediante un cuaderno de apuntes o con fotografías para luego transcribir a las hojas de Excel y quedar almacenadas para su posterior facturación y cobro por el consumo.

Tabla 2. Fallos que poseen en la gestión de información.

OPCIÓN	FRECUENCIA	%
Error de anotación	3	42,8
Pérdida de lecturas	2	28,6
Demora en buscar socios	2	28,6

Nos deja en evidencia que el proceso manual que realizan en la Junta presenta algunos inconvenientes resaltando los más relevantes como la demora en buscar los datos de un socio para verificar y facturar del consumo de agua, en otras ocasiones han redactado mal la información lo que provoca malestar con el usuario y por último por el cambio del personal se han llevado o no han entregado los informes haciendo que se desaparezca los datos de las lecturas del listado general, de acuerdo a los resultados obtenidos como de modela en la tabla 3es necesario implementar un sistema informático.

Tabla 3. Disponer de una aplicación móvil para mejorar el flujo de datos.

OPCIÓN	FRECUENCIA	%
SI	6	85.7
NO	1	14.3

Se puede deducir que la mayor parte del personal desea que exista un software informático en la Junta para de recolección de lecturas del consumo de agua de los usuarios y así mejorar los procesos tiempos de respuestas en la verificación de clientes, facturas mensuales cancelados, su ubicación geográfica, y brindar un servicio más adecuado y rápido a los socios. En el Diagrama de flujo de la figura 2. se visualiza el proceso manual de la información del consumo de agua que gestiona los administradores de la junta.

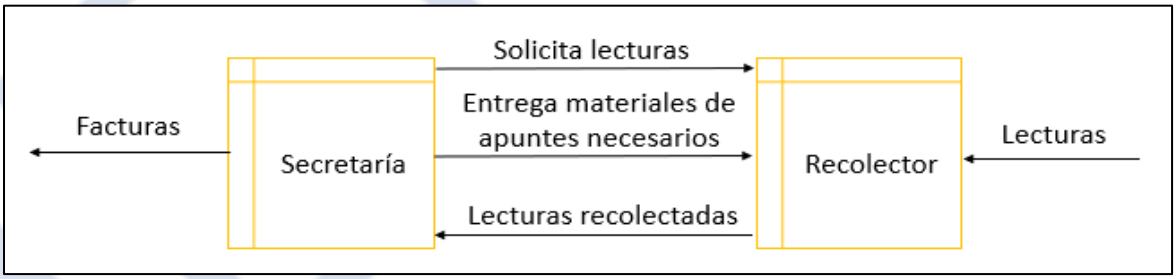

Figura 2. Flujo de datos de lecturas de medidores de los socios.

Diseño y Arquitectura de la Aplicación Móvil

El uso constante de celular hace más fácil realizar cualquier tipo de actividad con mayor comodidad, por los consiguiente se encuentra en un auge de desarrollo de sistemas orientadas a aplicaciones móviles que reemplazan a los sistemas de escritorio. Conforme a lo acordado, el equipo de trabajo está en plena condiciones de llevar a cabo el desarrollo de la aplicación móvil planeada por la junta de agua potable de la parroquia Santa Fe y poder dar solución a cualquier tipo de inconveniente que se puede presentar, en la tabla 4 describimos las actividades de cada tipo de usuario.

Tabla 4. Características de los usuarios para la aplicación móvil.

USUARIO	FUNCIÓN
OPERADOR	Es el encargado de tomar las lecturas de consumo, visualizar el listado de los socios, registrar y ver la ubicación de medidores, registrar los datos del consumo mensual del agua y generar reportes de lecturas.
SECRETARIA	Encargado de visualizar el listado de socios, ver ubicación de los medidores de los socios, generar reportes de lecturas.

Para el diseño de la navegación se ha centrado en una navegación lineal dando todas las facilidades al usuario de mantener una aplicación fácil de navegar, el inicio de sesión es para todos los usuarios en la cual deben ingresar un usuario y una contraseña asignado por el administrador del sistema. La navegación de interfaz para el operador y para la secretaria se muestra a continuación en la figura 3 y 4, cada uno de ellos con asignación específica de funciones:

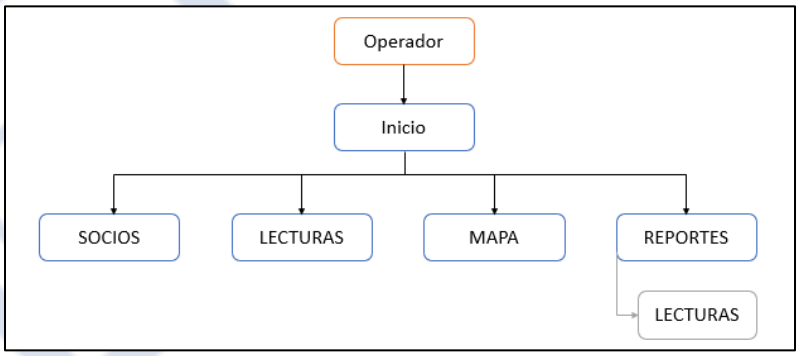

Figura 3. Navegación de interfaz para el Operador.

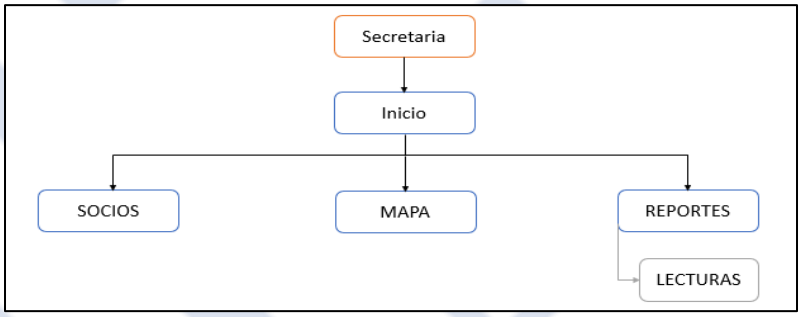

Figura 4. Navegación de interfaz para secretaria.

La aplicación UPIANA YAKU APP es de contenido dinámico con conexión a una base de datos (MySQL) alojado en un servidor, se utilizó el patrón de diseño modelo vista controlador (MVC):

Modelo. - Accede a la capa de almacenamiento de datos que maneja el sistema, su lógica de negocio y sus mecanismos de persistencia. Gestiona el acceso a dicha información en base a los privilegios de cada usuario.

Vista. - También conocida como Interfaz de Usuario, que presenta la información para su debida interacción.

Controlador. - Actúa como intermediario entre el Modelo y Vista, gestionando el flujo de información entre ellos y las transformaciones para adaptar los datos a las necesidades de cada uno.

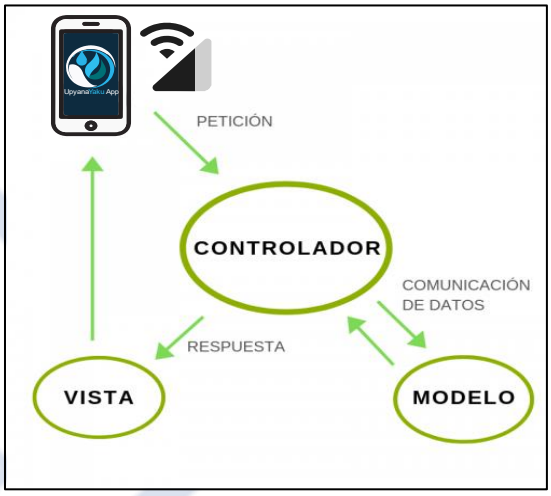

Figura 5. Patrón de diseño de la aplicación móvil.

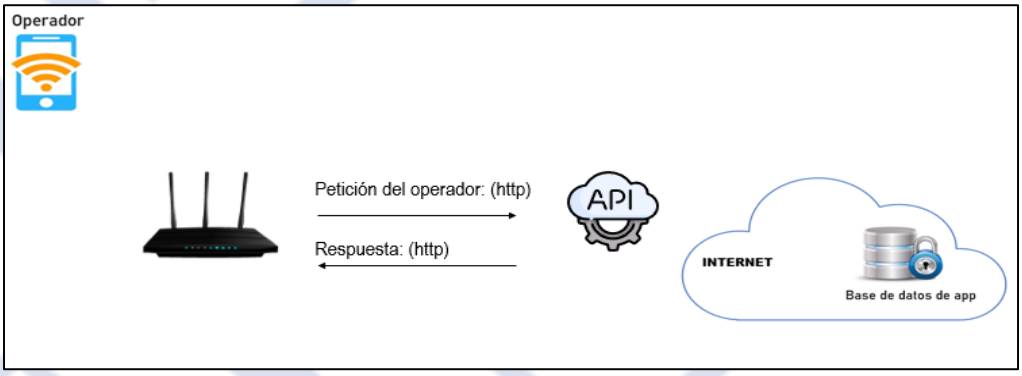

Figura 6. Arquitectura de la aplicación móvil.

Desarrollo de la Aplicación móvil

Para el desarrollo se utilizó las herramientas tecnológicas como una Computadora portátil, un teléfono móvil de la marca Samsung Galaxy J1 (LTE DUOS), el Software Visual Studio Code, Editor de código Sublime text 3 y el programa Día (para modelar diagramas); para el hosting PHP 7.1, MySQL 8.1 y Apache 2.5.

La aplicación Upiana Yaku App (Agua Potable Santa Fe) se ha centrado en el usuario, diseñando una aplicación enfocada en la interacción hombre – Celular, y se muestra un prototipo funcional probado en Android 10.0, en la Ventana de Splash Screen se visualiza la primera pantalla visible para el usuario cuando se inicia la aplicación. La pantalla de bienvenida es una de las pantallas más vitales de la aplicación, ya que es la primera experiencia del usuario con la App móvil.

Una vez desarrollado el sistema la primera ventana se muestra como indica la figura 7, que es la principal para que los usuarios encargados de la recolección de la toma de las lecturas del sistema de agua potable accedan a la misma con su nombre de usuario y contraseña, en el sistema se podrán visualizar los socios, rutas, medidores, ubicación y opción de ayuda (manual de navegación) para la recolección de información.

Figura 7. Ventana grafica de autentificación del usuario.

Posterior al ingreso a la aplicación se muestra la siguiente ventana que es la de menú principal, se muestra las diferentes opciones como detalla la figura 8, que tendrá el usuario para que pueda ingresar y cumplir con las funciones asignadas de tal manera que se le facilitara hacer su trabajo mediante la aplicación móvil que fue diseñada.

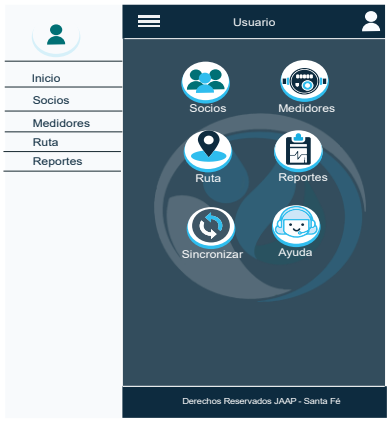

Figura 8. Ventana grafica del menú principal.

En la figura 9, se muestra la distribución de los sub menús como: ***Socios*** el usuario podrá ingresar, actualizar y verificar la información existente de todos los clientes en el sistema; en el sub menú ***Medidores*** se realizará el nuevo ingreso, actualización y verificación de los datos del medidor de cada socio; en el sub menú ***Ruta*** se observa sobre un mapa la ubicación de cada medidor de agua registrado e indica que medidores también faltan por registrar las lecturas, así como la ruta a seguir en caso de desconocer su paradero; sub menú ***Reportes*** se visualiza toda la información asociada entre socio-medidor-ubicación la cual puede ser descargada e impresa en hoja de papel para su facturación; el sub menú ***Sincronización*** es él envió de los datos en pequeños paquetes con la finalidad de que tanto el emisor como el receptor tengan la misma información, para este proceso los registros de lecturas de agua potable deben estar guardados en el dispositivo móvil, una vez que se enlace una conexión mediante datos móviles 3G o red Wifi, con tan solo presionar el icono de sincronización; y la opción ***Ayuda*** esta un manual de usuario en donde indica de cómo se debe usar el sistema.

A continuación, se muestra las diferentes opciones del sub menú de la aplicación:

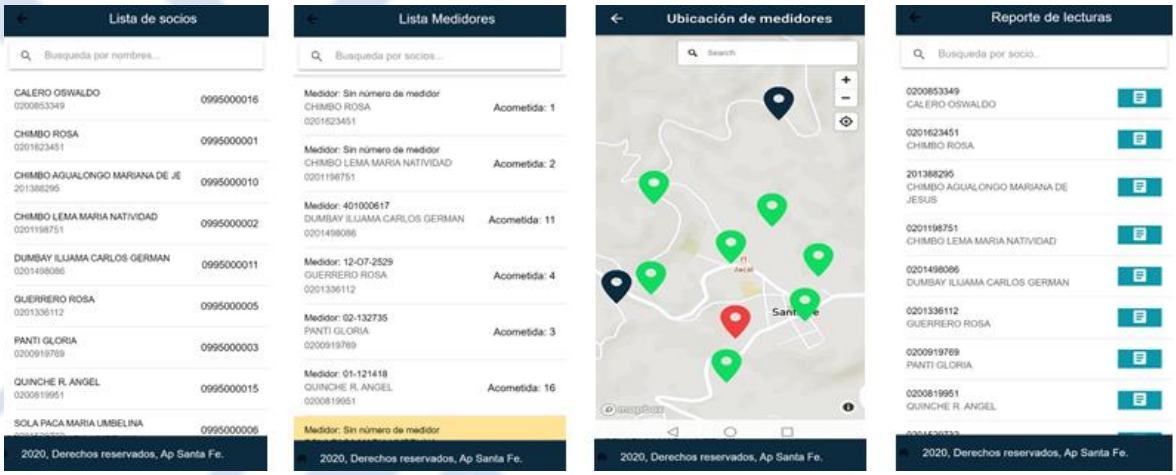

Figura 9. Ventana grafica de los sub menús del usuario en la aplicación móvil.

Una vez culminado el desarrollo de la aplicación web se ha podido determinar lo siguiente:

El **módulo para la recolección de lecturas** cuenta con la asignación de trabajos de inspección, las acometidas están geográficamente localizadas en donde la aplicación móvil traza la ruta para recabar los datos, la directiva de la junta verifica sí acudió al lugar asignado mediante puntos de referencias, se podrá visualizar la ubicación si anteriormente tomo los puntos geográficos de cada acometida, el cierre del ciclo de tomas de lecturas se verifica cuando a uno o varios clientes no se ha registrado y mostrara un mensaje de advertencia, el responsable de las lecturas puede añadir observaciones sobre el estado de las acometidas, la planilla evidencia el valor real que el cliente debe cancelar por el servicio, al registrar la lectura se genera una planilla de consumo del agua y se ejecuta un algoritmo en la cual verifica si existen otros valores adicionales desarrollados en otros módulos que no refleja la presente propuesta tales como instalación, reparaciones, convenios de pagos, y multas.

El **módulo para los puntos geográficos** se requiere que el teléfono este activo el GPS para poder marcar las coordenadas y se requiere una conexión a una red de datos móviles o wifi, o a su vez tener descargado los mapas del lugar donde va a marcar dichos puntos en modo offline. A continuación, en la figura 10 se observa el flujo de información del sistema:

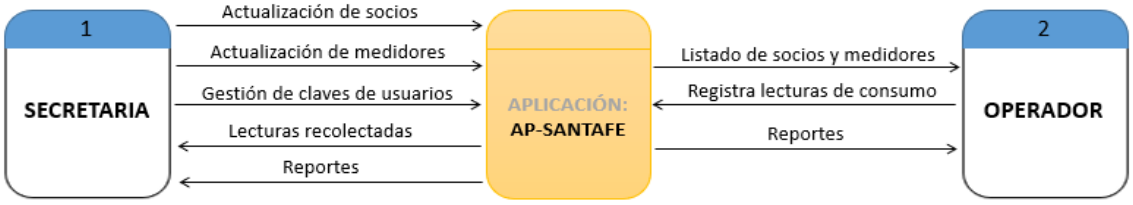

Figura 10. Flujo de datos de la aplicación móvil (Uypaya yaku App).

Aplicación del plan de pruebas

El propósito de realizar las pruebas de verificación y validación (modelo V) a la aplicación móvil como se muestra la distribución de procesos en la figura 11, fue para obtener un producto de calidad al finalizar proceso de desarrollo, todo en base a los estándares y requerimientos de los usuarios del software, durante el proceso de construcción se realizó pruebas en simuladores virtuales como en dispositivos físicos con sistema operativo Android versión 5.1 en adelante para verificar su funcionalidad, y la validación fue en conjunto entre los desarrolladores con los usuarios finales con el objetivo de detectar fallos o errores y realizar las correcciones necesarias.

Para verificar que el funcionamiento de la aplicación móvil para la toma de lecturas está operando correctamente fue necesario hacer la simulación de ingreso de datos mediante un servidor de

pruebas, con el objetivo de identificar errores que podría presentar la aplicación móvil y de tal manera corregir los mismos antes de implementar el proceso de operación y puesta en marcha en la institución.

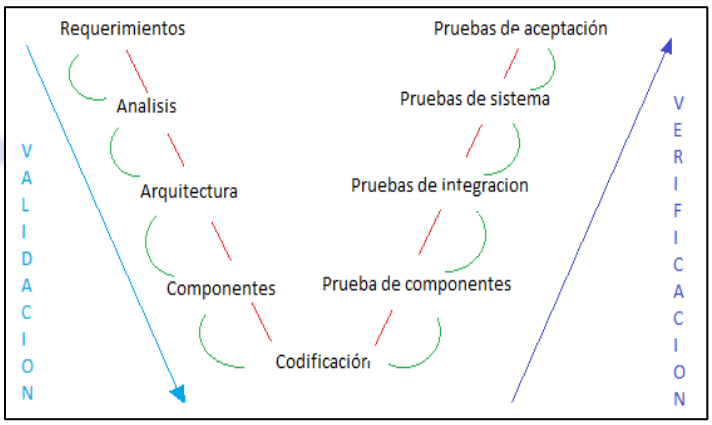

Figura 11. Modelo de prueba en V ejecutado a la aplicación móvil.

Las pruebas que se realizaron con el fin de encontrar errores en su funcionalidad en la aplicación web (en desarrollo por otro grupo) y móvil, tomando como referencia los resultados esperados se tomaron en cuenta los siguientes casos de prueba como tipo, los objetivos y el cumplimiento que detallan en la tabla 5 y sus características más importantes con la tabla 6.

Tabla 5. Ejecución del plan de prueba (modelo V).

Tipo	Objetivo	Cumplimiento
Componentes	Verificar que los trozos de código implementado para el desarrollo de la aplicación se ejecuten sin errores.	SI
Integración	Comprobar que el código de dos o más módulos se fusione entre sí y a su vez que interactúen con las interfaces.	SI
Sistema	Asegurar la apropiada navegación (experiencia de usuario) dentro de la aplicación, ingresando datos reales de acuerdo a los requerimientos de los usuarios que se encuentran en las Historias de la documentación.	SI
Aceptación	Garantizar que la aplicación móvil cumpla con los requisitos establecidos y satisfaga las necesidades de los usuarios.	SI
Funcionalidad	Cotejar el posicionamiento de las rutas (mapa-ubicación geográfico) y su	SI

	recuperación de los datos del consumo del Web Services del sistema web. Contrastar la información de la toma de lecturas de los usuarios y su generación de planillas.	
Seguridad	Aplicar las normas de seguridad tanto en el software y hardware con una configuración mínima requerida y evitar vulnerabilidades de algún ataque malicioso.	SI
Integridad	Evitar la duplicidad de datos cuando un cliente tenga dos o más medidores bajo su responsabilidad.	SI

Tabla 6. Características más relevantes de la aplicación móvil

N°	Interrogantes	Respuestas	Justificación
1	Diseño final para el usuario	La interfaz será interactiva y amigable para el majeo	La aplicación se desarrolló para dispositivos ANDROID.
2	Actualización de datos	La información se actualizará On-line.	La información será manejada con una base de datos de manera On-line.
3	Nivel de complejidad	El proceso será muy complejo para la salida/entrada.	El usuario de la aplicación móvil realizara múltiples actividades.
4	Aplicación reutilizable	La aplicación será reutilizable para otros proyectos	Sí, con los sistemas desarrollados para ANDROID.
5	Rendimiento	No es un obstáculo para la aplicación.	La aplicación no cuenta con herramientas adicionales.
6	Se utilizará configuraciones complejas	Ninguna restricción que limite la utilización de la aplicación.	No se utilizará ninguna configuración para la aplicación.
7	Instalación rápida	Fue establecida por el usuario requerimientos	La instalación se podrá realizar en un dispositivo

		específicos, y será fácil de instalar	ANDROID con versiones superiores a 5.1.
8	Instalación en múltiples dispositivos	La aplicación será instalada solo en dispositivos con ANDROID.	No se ha realizado para otro tipo de sistemas operativos.
9	Información de datos	Será por medio de protocolos de comunicación.	La aplicación podrá realizar la comunicación por internet.

Discusión

En la presente investigación se busca conocer que, al desarrollar una aplicación móvil para la administración y gestión del consumo de agua potable, SI mejoraran o NO los procesos de recolección, búsqueda, visualización y actualización de datos de consumo del agua potable en la Junta Administradora de la parroquia Santa Fe.

La aplicación móvil implementado para la JAAP optimiza los procedimientos en la gestión administrativa del registro de consumo de agua, automatizando los procesos de registro, actualización de lecturas de medidor y la generación de planillas para los pagos que efectúan los usuarios, y la visualización de su ubicación geográfica para una localización oportuna, así da un mayor sustento al estudio con otro proyecto similar de [27] en cual analiza el proceso de registro de lecturas del medidor mediante una aplicación móvil y "se determinó que existe un reducción del 49,25% en relación al tiempo que se emplea en realizar el mismo proceso manualmente".

La tecnología móvil cada si se va adaptando a las necesidades de la sociedad y va convirtiendo en una herramienta de trabajo con su uso diario, y se va dejando de utilizar los polígrafos, cuadernos, folder, mapas tradicionales y los programas tradicionales de una computadora de escritorio o laptop y están siendo reemplazado rápidamente por las aplicaciones móviles los cuales facilitan su manejo con tiempos de respuestas instantáneas, y también ya se han introducido en los sistemas de agua potable ayudando a llevar un control adecuado de los registros de socios y cuentas financieras, también otra investigación similar de [28] manifiesta que "se reemplaza las hojas de lectura y el polígrafo que utiliza el personal de la empresa para el registro manual por un registro online" el cual permite tener de inmediato la información que está almacenado en la base de datos, así como el estudio de [29] las herramientas tecnológicas ayuda a "mejorar la

eficiencia de la administración del sistema de cobro de agua potable" con la actualización de los sistemas manuales hacia una utilización en una APP móvil para la recolectar la información.

El desarrollo del proyecto fue brindar mejoras a los procesos de registro de clientes, tomas de lecturas, generación de reportes, efectuar facturas cobro, ubicación geográfica mediante una aplicación móvil para una gestión directa y fácil, y para la sincronización de datos tendrá que contar con conexión a internet pudiendo acceder desde cualquier lugar, hoy en día se requiere de agilidad y disponibilidad inmediato de los datos, así como nos muestra el estudio de [30] con la APP "permite realizar la toma de lecturas a los domicilios reduciendo el tiempo de registro en un 97,33% frente al registro manual".

Conclusiones

En la construcción del proyecto se recolecta la información mediante la aplicación de las técnicas e instrumentos y herramientas de investigación, y logra determinar la situación actual de la problemática, se utiliza métodos y técnicas de desarrollo de software como la Scrum-Ágil el mismo que facilita la gestión de la información, interfaz amigable y facilidad de uso, además de aplicar un plan de pruebas (modelo V) para su evaluación y garantizar la calidad del producto.

Con el sistema informático antes mencionado se mejora los procesos del flujo de información de las lecturas y socios de la Junta Administradora de Agua Potable de la Parroquia Santa Fe, mediante la aplicación móvil para los teléfonos inteligentes se actualiza el sistema obsoleto, desorganizado y desactualizado de cobros que mantenía en la JAAP y se reemplaza las hojas de papel y el polígrafo por un registro off-line u on-line que se puede realizar directamente el software informático, por ende su uso ayuda a tener una mejor administración y gestión de cobro del consumo de agua en los domicilios, garantizando la disponibilidad de los datos en cualquier momento y brindando una mejor atención al usuario.

Los resultados obtenidos en el estudio nos muestran que la implementación fue satisfactoria, los participantes con respecto a la usabilidad y experiencia en el manejo de la aplicación móvil para la toma de lecturas, generación de planillas, sondeo de información expresaron su satisfacción con la navegación e iteración, y el software funciona utilizando un dominio y hosting para que los datos sean sincronizados y almacenados en una base de datos, a los cuales el personal administrativo de la JAAP pueden acceder desde sus teléfonos o tablets que tenga un sistema operativo igual o superior Android 5.0 para su instalación y pueden buscar, visualizar, actualizar la información.

Referencias

[1] J. H. Bustos Parra y H. A. Anillo Castellar, «Wireless application protocol WAP.,» 2004. [En línea]. Available: http://biblioteca.utb.edu.co/notas/tesis/0031332.pdf. [Último acceso: 2021].

[2] A. U. Gawas, «An overview on evolution of mobile wireless communication networks: 1G-6G,» *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 3, nº 5, pp. 3130-3133, 2015.

[3] A. Baz Alonso, I. Ferreira Artime, M. Álvarez Rodríguez y R. García Baniello, «Dispositivos móviles,» *EPSIG Ing. Telecomunicación,* vol. 12.

[4] G. de Lucas, «EVOLUCION DE LAS APLICACIONES PARA MOVILES,» 2017.

[5] F. Sáez y J. Adell, «De los ordenadores a los dispositivos móviles,» Barcelona, Graó, 2015, pp. 11-29.

[6] M. E. Alonso del Barrio y M. Antón Crespo, «Los contenidos periodísticos en los medios para dispositivos móviles: la adaptación a la evolución tecnológica.,» Madrid, 2016.

[7] J. M. Aguado, I. J. Martínez y L. Cañete-Sanz, «Tendencias evolutivas del contenido digital en aplicaciones móviles.,» *Profesional de la Información,* pp. 787-796, 2015.

[8] I. Márquez, «El smartphone como metamedio,» vol. 11, nº 2, pp. 061-071 , 2017.

[9] G. M. M. GUADALUPE, «USOS Y TIPOS DE APLICACIONES MÓVILES,» Oaxaca, 2015.

[10] H. A. Botello-Peñaloza, «Determinantes del acceso al internet: Evidencia de los hogares del Ecuador.,» *Entramado,* vol. 11, nº 2, pp. 12-19, 2015.

[11] J. P. D. A. Ponce, «Ecuador Estado Digital,» Quito-Ecuador, 2021.

[12] M. d. T. y. d. l. S. d. l. I. MINTEL, «Ecuador Digital,» 2020. [En línea]. Available: https://www.telecomunicaciones.gob.ec/25693-2/. [Último acceso: 2021].

[13] A. D. L. Espriella-Babiloni, «Comparación entre tecnologías emergentes y tradicionales en automatización e instrumentación industrial.,» *Sostenibilidad, Tecnología y Humanismo,* vol. 10, nº 1, pp. 70-77, 2019.

[14] B. O. C. GONZÁLEZ, «ACEPTACIÓN DEL USO DE LA TECNOLOGÍA MÓVIL EN EL PAGO DE SERVICIOS BÁSICOS–AGUA, TELÉFONO Y ENERGÍA ELÉCTRICA,» Guatemala, 2017.

[15] J. C. Nilcia, H. L. Adriana María, E. D. Ricardo y Z. I. Nadiova Victoria, «El agua: recurso vital para la supervivencia humana.».

[16] C. Chulluncuy y C. Nadia, «Tratamiento de agua para consumo humano,» *Ingeniería Industrial,* vol. 029, pp. 153-170, 2011.

[17] N. MARÍA VIRGINIA y B. S. HENRY A., «Estrategias de mejora continua en plantas potabilizadoras Venezolanas,» *Revista de la Facultad de Ingeniería Universidad Central de Venezuela,* vol. 29, nº 1, pp. 37-50, 2014.

[18] M. L. Ramos Bayas, «El capital social de Juntas Administradoras de Agua Potable y Riego del Ecuador JAAPRE y la Ley Orgánica de recursos hídricos, usos y aprovechamiento del agua (2009–2015),» Ecuador: Flacso Ecuador, Quito, Ecuador, 2017.

[19] Z. Stefan, V. V. Wagner y R. C. Hermes, «La seguridad de medidores de agua potable contra robo, vandalización y manipulación–problemática, avances y propuesta,» *INDES Revista de Investigación para el Desarrollo Sustentable,* vol. 3, nº 2, pp. 5-15, 2017.

[20] M. E. PETIT-BREUILH SEPÚLVEDA, Desastres naturales y ocupación del territorio en Hispanoamérica, vol. 70, Servicio de Publicaciones de la Universidad de Huelva., 2018.

[21] J. A. Mera Paz, «Análisis del proceso de pruebas de calidad de software,» *Ingeniería solidaria,* vol. 12, nº 20, pp. 163-176, 2016.

[22] R. Hernández-Sampieri y C. P. M. Torres, Metodología de la investigación (Vol. 4), México^ eD. F DF: McGraw-Hill Interamericana., 2018.

[23] G. P. J. Antonio, Técnicas e instrumentos para la recogida de información, Editorial UNED, 2016.

[24] T. D. D. L. NEFTALI, «Población y muestra,» 2016.

[25] T. Dimes, Conceptos Básicos de Scrum: Desarrollo de software Agile y manejo de proyectos Agile, Babelcube Inc., 2015.

[26] M. C. Gasca Mantilla, L. L. Camargo Ariza y B. Medina Delgado, «Metodología para el desarrollo de aplicaciones móviles,» *Tecnura,* vol. 18, nº 40, pp. 20-35, 2013.

[27] T. A. Tisalema P., «Desarrollo de una aplicación web/móvil para el registro de consumo/pago de los usuarios de la Junta Administradora de Agua Potable Angahuana Alto, aplicando TDD,» Riobamba, 2019.

[28] J. C. Tapia y J. O. Castro, «Ingreso de lecturas de consumo de agua potable en EMAPAL-Azogues, a través de dispositivos móviles,» *Revista Científica y Tecnológica UPSE,* vol. 4, nº 2, pp. 85-90, 2017.

[29] Y. M. ESCORZA-SÁNCHEZ, C. ALAMILLA-CINTORA, G. MARTÍNEZ-MARTÍN y Y. SALDAÑA-TAPIA, «Herramienta para la administración del sistema de cobro de agua potable.,» *Revista de Tecnología Informática,* vol. 1, nº 1, pp. 36-45, 2017.

[30] B. A. Ramírez Rodríguez, «Implementación del sistema de gestión informático para la reducción de pérdidas de consumo en el sistema de agua potable de la Junta Administradora de Agua Potable-Zapotal. módulo: toma de lectura y generación de planillas,» La Libertad, 2019.

Aplicación de los árboles de decisión en el diagnóstico de Anemia en niños de la ciudad de Arequipa

Application of Decision Trees in the diagnosis of Anemia in children in the city of Arequipa

Indira Agramonte Mayhua
Universidad Nacional de San Agustín.
Arequipa, Perú.
iagramontem@unsa.edu.pe

Alex Chaco Huamani
Universidad Nacional de San Agustín.
Arequipa, Perú.
achacohu@unsa.edu.pe

Alexander Valdiviezo Tovar
Universidad Nacional de San Agustín.
Arequipa, Perú.
avaldiviezot@unsa.edu.pe

Melody Ramos Challa
Universidad Nacional de San Agustín.
Arequipa, Perú.
mramoschal@unsa.edu.pe

ARK: ark:/42411/s9/a69
PURL: 42411/s9/a69

RECIBIDO 22/05/2022 • ACEPTADO 14/06/2022 • PUBLICADO 30/09/2022

RESUMEN

Uno de los problemas más comunes en los niños que no son correctamente alimentados es la anemia. La deficiencia de hierro es perjudicial para los menores, pues impide que realicen sus actividades diarias por el cansancio extremo y fatiga. Debido a esta situación, el Estado peruano ha intentado disminuir el nivel de prevalencia de anemia a nivel nacional con campañas médicas en diferentes regiones, pese a ello, localidades como Caylloma en Arequipa aún mantienen un alto porcentaje de infantes anémicos, para ello se desarrolló una implementación mediante Árboles de Decisión con el lenguaje Python para poder determinar si un niño tiene anemia en base a los datos proporcionados.

Palabras claves: Árbol de decisión, anemia, inteligencia artificial.

ABSTRACT

Anemia is one of the most common problems in children who are not adequately fed. Iron deficiency harms minors as it prevents them from carrying out their daily activities due to extreme

tiredness and fatigue. Due to this situation, the Peruvian State has tried to reduce the level of prevalence of anemia at the national level with medical campaigns in different regions; despite this, localities such as Caylloma in Arequipa still maintain a high percentage of anemic infants, for which an implementation through Decision Trees with the python language to be able to determine if a child has anemia based on the data provided.

Keywords: *Decision tree, anemia, artificial intelligence.*

INTRODUCCIÓN

Uno de los mayores problemas de salud pública en el mundo asociado al incremento de índices de morbilidad y mortalidad, es la anemia [1]; según la OMS (Organización Mundial de la Salud) en el año 2020 la anemia afectó en todo el mundo a 1620 millones de personas, lo que representaría al 24,8% de la población, mientras que, la máxima prevalencia se da en los niños en edad preescolar con 47,4% [2].

En nuestro país muchas mujeres gestantes y niños hasta los 11 años de edad mueren debido a tardíos diagnósticos por parte de los médicos siendo familias del sector rural los más afectados, esto debido a deficiencias nutricionales, bajos ingresos familiares, bajo nivel educativo, políticas de salud centralistas y al bajo financiamiento destinado por parte del estado peruano [3].

Realizar diagnósticos en niños no es una tarea fácil, no sólo basta con saber si un niño presenta anemia o no, sino también conocer el grado de afección; dato que podría ser determinante en la evaluación de futuras complicaciones en el paciente [4]. Por otro lado, existen diferentes indicadores para realizar tales diagnósticos los más comunes son talla, peso, niveles de hemoglobina, frecuencia cardíaca entre otros lo que permite evaluar la condición en la que se encuentra un individuo frente a esta enfermedad [5]. Otro factor importante a considerar también es la ubicación geográfica, en nuestro país la media de niveles de hemoglobina en regiones de la sierra es mayor que en las regiones de la costa.

Tomando en cuenta lo anteriormente descrito, en este artículo se pretende desarrollar un Árbol de decisión, basado en un sistema de información del estado nutricional de niños del Perú proporcionada por el Instituto Nacional de Salud con el fin de conocer el diagnóstico de anemia en niños pertenecientes a la región de Arequipa, de esta manera también anticipar el nivel de afección con una precisión aceptable [6]. En ese sentido nos basamos en el Sistema de Información del Estado Nutricional (SIEN) que fue implementado por el Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición (INS/CENAN) [7] con el fin de obtener un diagnóstico de anemia en niños menores a 11 años en la ciudad de Arequipa ya que dicho sistema realiza un proceso continuo y sistemático que registra, procesa, reporta y analiza información del estado nutricional de niños menores de cinco años y mujeres gestantes que acuden a

establecimientos de salud del primer nivel de atención del Ministerio de Salud [8]. El Árbol de decisión presentado es una estructura que nos ayudará a tomar como entrada un objeto o situación descrita por un conjunto de propiedades en este caso indicadores de anemia y proporcionará como salida una decisión de sí o no. En términos de capacidad, el Árbol de decisión es un método rápido y eficaz que va a permitir clasificar las entradas del conjunto de datos y proporcionar una buena capacidad de apoyo a las decisiones [9].

El sistema será de gran ayuda en el diagnóstico de anemia ya que tiene ciertos parámetros que pueden ser revisados utilizando una estructura de inteligencia artificial para conseguir datos más cercanos que nos permitan encontrar una relación entre el nivel de hemoglobina frente a un posible cuadro de anemia [10].

Trabajos Relacionados

Barrientos et al. [11] evalúan el desempeño de tres algoritmos basados en los árboles de decisión a partir de los resultados en su aplicación con el fin de determinar si la técnica puede llegar a ser una herramienta de soporte y ayuda eficaz en el tratamiento y diagnóstico médico correspondientes a la sintomatología que un médico especialista considera importante en el diagnóstico de cáncer de seno.

Para realizar esto los autores utilizaron dos bases de datos las cuales contienen datos recopilados por expertos en patología con experiencia en la detección de cáncer de mama, estos datos permiten el diagnóstico para saber si el paciente tiene o no cáncer de mamá sin embargo es necesario una biopsia y una mamografía para indicar si el tumor es maligno. Esta investigación concluye que los árboles de decisión son posibles de construir a partir de datos médicos, sin embargo, los datos recopilados deben de pasar por un proceso de clasificación de esta manera se obtiene un margen de error mínimo. R. Yen et al.[12] examinaron los factores de riesgo principales relacionados con las conductas alimentarias inadaptadas y la regulación de las emociones, y cómo sus interacciones afectan a la detección de los trastornos alimentarios. Para ello construyeron un modelo de árbol de decisión utilizando el aprendizaje automático sobre un conjunto de datos de 830 mujeres jóvenes chinas sin antecedentes con una edad media de 18,91 años.

El conjunto de datos se dividió en datos de entrenamiento y de prueba con una proporción de 70% y 30%. Los resultados señalan que la inflexibilidad de la imagen corporal se identificó como el principal clasificador de las mujeres con alto riesgo de TCA, seguidas por la angustia psicológica y la insatisfacción corporal. Pei et al.[13] establecieron un modelo predictivo basado en los factores de riesgo asociados a la diabetes mediante un árbol de decisión. En su investigación reclutaron a un total de 10.436 participantes que se sometieron a un examen de salud entre enero de 2017 y julio de 2017, de los cuales 3454 permanecieron en el conjunto final de datos.

Para asignar el porcentaje de datos de entrenamiento y de prueba usaron el algoritmo J48, dando como resultado un 70% de datos para entrenamiento y 30% de datos de prueba. Su trabajo consta de 14 variables de entrada y 2 de salida, el modelo de árbol de decisión presentado identificó varios factores clave que están estrechamente relacionados con el desarrollo de la diabetes y que también son modificables. Además, el modelo alcanzó una exactitud de clasificación del 90,3% con una precisión del 89,7% y un recuerdo del 90,3%.

Presentando como conclusión que su modelo de árbol de decisión estima el desarrollo de la diabetes en una población china adulta de alto riesgo con un gran potencial para la aplicación del control de la diabetes. Bouza y Santiago [14] califican a los árboles de decisión como una herramienta muy potente, que permite la segmentación, clasificación, predicción, reducción, identificación y recodificación de los datos. En el campo de la medicina la toma de decisiones es fundamental, es por esto que algunas instituciones utilizan sistemas conocidos como Decision support systems, sistemas eficientes y fiables que ayudan a los médicos a obtener sus diagnósticos. Tomando en cuenta las investigaciones anteriormente descritas, se plantea realizar un árbol de decisión que pueda obtener un diagnóstico de anemia en base a un conjunto de datos que pueden ser revisados utilizando una estructura de inteligencia artificial.

Marco Teórico

A. Árboles de Decisión

Según Solarte y Soto [15] los árboles de decisión es una técnica de aprendizaje inductivo supervisado no paramétrico, se utiliza en la predicción y se emplea en el campo de la inteligencia artificial. Por otro lado, Landín y Romero [16] nos indican que los árboles de decisión posibilitan la uniformidad del diagnóstico y tratamiento en pacientes. Se basa en la aplicación de un conjunto de reglas y el uso de funciones lógicas, un aporte del uso de esta técnica es el plantear un problema de forma concisa y que a partir del mismo problema analizar todas las opciones posibles [17].

B. Data reduction (python)

Data reduction es una técnica para reducir la cantidad de variables en un conjunto de datos[18][19]. En tareas de Machine Learning como la regresión o la clasificación, a menudo hay demasiadas variables, conocidas también como características, con las que trabajar. Cuanto mayor sea el número de características, más difícil será modelarlas. Además, algunas de estas, pueden ser bastante redundantes, añadiendo ruido al conjunto de datos y no tiene sentido tenerlas en los datos de entrenamiento. Es aquí donde se tiene que reducir la cantidad de variables[19].

El proceso de data reduction transforma los datos de un conjunto de características de alta dimensión a un conjunto de características de baja dimensión a través de dos componentes: la selección de características y la extracción de características. En la selección de características, se eligen subconjuntos más pequeños de características de un conjunto de datos de muchas dimensiones para representar el modelo mediante el filtrado, la envoltura o la incrustación. La extracción de características reduce el número de dimensiones de un conjunto de datos para modelar las variables y realizar el análisis de componentes. También es importante que las propiedades significativas presentes en los datos no se pierdan durante la transformación [18][19].

C. Análisis de datos

Histogramas:

En la presente sección se mostrará los histogramas pertenecientes a algunas variables a utilizar en el presente estudio. Estos diagramas ayudan a comprender la distribución de los datos y comprobar valores extremos o atípicos.

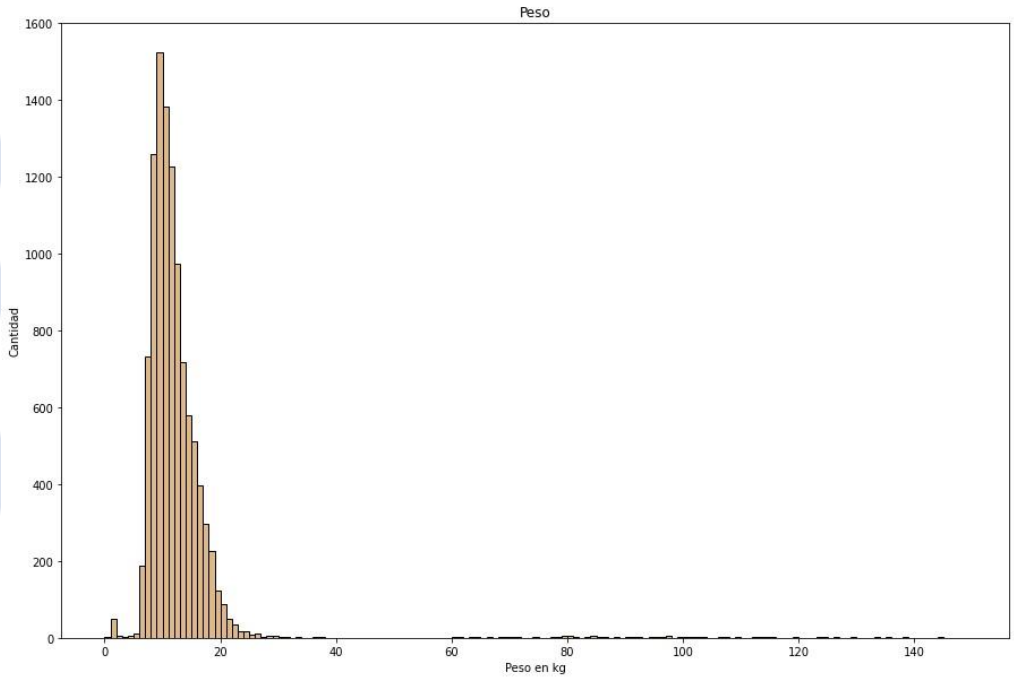

Figura 1. Cantidad de niños/niñas divididos por peso.

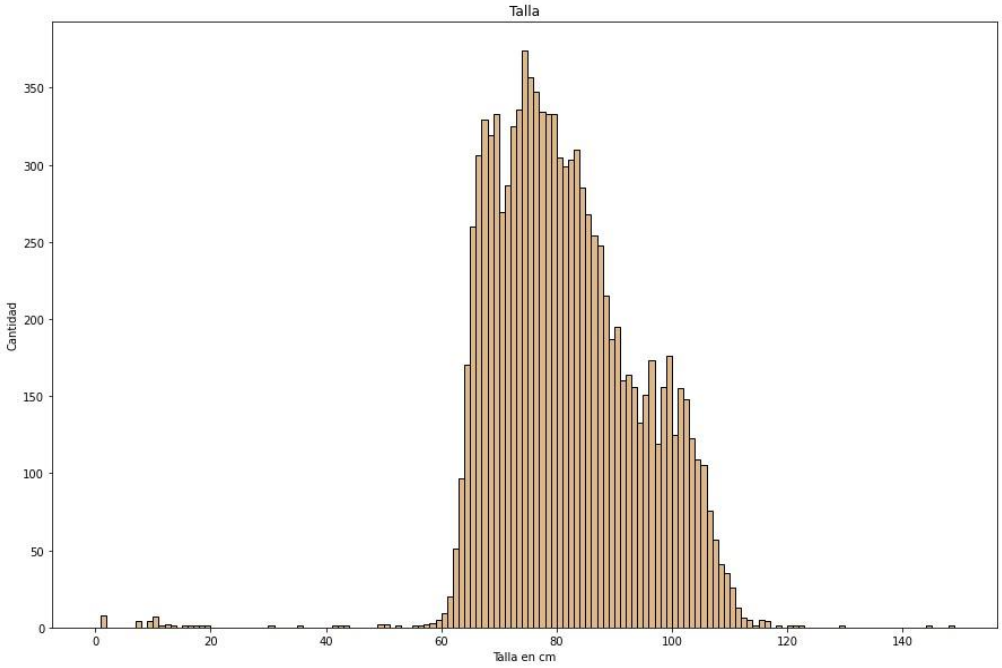

Figura 2. Cantidad de niños/niñas agrupados por la estatura.

Metodología y Desarrollo

A. Recursos

Para realizar esta investigación fue necesaria la búsqueda de datos que nos permitieran tener acceso a información verídica. Por esta razón el dataset utilizado fue extraído del Repositorio de Datos del Instituto Nacional de Salud, la investigación que lo acompaña es Sistema de Información del Estado Nutricional de niños y gestantes Perú - INS/CENAN. Las herramientas que se utilizaron para realizar este proyecto fueron: Visual Studio Code, Google Colab que nos permiten crear código para data reducción y crear el árbol de decisión a partir de un data set.

B. Limpieza de datos

Data Reduction, es el proceso de reducir la cantidad de capacidad requerida para almacenar datos. La reducción de datos puede aumentar la eficiencia del almacenamiento y reducir los costos. Los proveedores de almacenamiento a menudo describirán la capacidad de almacenamiento en términos de capacidad bruta y capacidad efectiva, que se refiere a los datos después de la reducción. En esta ocasión lo que se busca es reducir la cantidad de

atributos(columnas) del dataset Niños Arequipa, esta reducción es debido a que mucha de esta información es irrelevante y en ocasiones inutil.

```
if row['Peso'] == '#¡NULO!' or row['Talla'] == '#¡NULO!':
        continue
```

Con la anterior sentencia se elimina o se evita aquellas filas con datos incompletos, es decir, aquellas filas que tengan como contenido !#NULO! . Esto se hace para que esta información no se tome en cuenta por que esta información se vería reflejada en el árbol de decisión que se realizará a posterior.

```
Diresa = row['Diresa']
DxAnemia = row['Dx_Anemia']
Sexo = row['Sexo']
EdadMeses = row['EdadMeses']
Peso = row['Peso']
Talla = row['Talla'] Hemoglobina =
row['Hemoglobina']
HBC = row['HBC']
ProvinciaREN = row['ProvinciaREN']
DistritoREN = row['DistritoREN']
```

Para la siguiente parte, el código nos muestra un filtro donde solo se escogen los atributos que son importantes. En otras palabras las atributos con una mayor relevancia son las elegidas para que continúen en un nuevo archivo csv, además DxAnemia es nuestro dato de salida por lo que también se le toma en cuenta.

```
with open('Niños_AREQUIPA_V2.csv', 'w', newline='') as csvfile:
        fieldnames = ['Diresa', 'DxAnemia','Sexo', 'EdadMeses',
'Peso','Talla','Hemoglobina','HBC','ProvinciaREN','DistritoREN']
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
    writer.writeheader() writer.writerows(new_data)
```

Finalmente se crea un nuevo archivo csv con la información separada. Este nuevo archivo csv llamado “Niños_AREQUIPA_V2.csv” contiene una cabezera con los siguientes datos: 'Diresa', 'DxAnemia', 'Sexo', 'EdadMeses', 'Peso', 'Talla', 'Hemoglobina', 'HBC', 'ProvinciaREN', 'DistritoREN'. y es acompañado por la información que está separado en un array new_data.

Diresa	DxAnemia	Sexo	EdadMeses	Peso	Talla	Hemoglobina	HBC
AREQUIPA	Anemia Leve	F	52	16.90	109.0	11.8	10.67
AREQUIPA	Normal	M	14	11.09	82.1	13.8	13.47
AREQUIPA	Normal	M	10	10.67	73.3	11.8	11.47
AREQUIPA	Normal	F	13	9.22	74.8	12.3	11.97
AREQUIPA	Normal	F	14	8.51	73.7	11.5	11.49
...	...	...	...	...	...	...	...
AREQUIPA	Normal	F	48	19.60	109.8	13.5	12.37
AREQUIPA	Normal	F	6	7.20	72.0	13.8	12.67
AREQUIPA	Normal	M	6	7.50	67.0	12.9	11.77
AREQUIPA	Normal	M	24	12.90	85.0	14.7	13.64
AREQUIPA	Normal	M	46	19.00	102.5	15.2	14.14

Diresa	ProvinciaREN	DistritoREN
AREQUIPA	AREQUIPA	PAUCARPATA
AREQUIPA	CAYLLOMA	MAJES
AREQUIPA	CAYLLOMA	MAJES
AREQUIPA	CAYLLOMA	MAJES
AREQUIPA	CONDESUYOS	RIO GRANDE
...	...	...
AREQUIPA	AREQUIPA	PAUCARPATA
AREQUIPA	AREQUIPA	PAUCARPATA
AREQUIPA	AREQUIPA	CERRO COLORADO
AREQUIPA	AREQUIPA	MARIANO MELGAR
AREQUIPA	AREQUIPA	MARIANO MELGAR

[10553 rows x 9 columns]

Figura 3. Limpieza de datos

C. Librerías a utilizar

Para elaborar Árbol de Decisión en python se utilizaron las siguientes librerías:

Pandas: Pandas es una librería de Python especializada en el manejo y análisis de estructuras de datos. Las principales características de esta librería son: Define nuevas estructuras de datos basadas en los arrays de la librería NumPy pero con nuevas funcionalidades. Permite leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL. Permite acceder a los datos mediante índices o nombres para filas y columnas. Ofrece métodos para ordenar, dividir y combinar conjuntos de datos. Permite trabajar con series temporales.

DecisicionTreeClassifier : La biblioteca Scikit Learn tiene una función de módulo DecisionTreeClassifier() para implementar el clasificador de árboles de decisión con bastante facilidad.

Confusion_matrix : La matriz de confusión sirve para evaluar la precisión de una clasificación. Por definición, una matriz de confusiones tal que es igual al número de observaciones que se sabe que están en el grupo y se predijo que estén en el grupo.

StringIO : Proporciona un medio conveniente para trabajar con texto en memoria utilizando la interfaz de programación de archivo (read() , write() , etc.). Utilizamos StringIO para construir cadenas grandes puede ofrecer ahorros en el rendimiento sobre algunas otras técnicas de concatenación de cuerdas en algunos casos.

Pydotplus : Se pueden utilizar tanto para resolver problemas de clasificación como de regresión. Una de sus principales ventajas es la facilidad con la que se puede interpretar los resultados en base a reglas. Permitiendo no solo obtener un resultado, sino inspeccionar los motivos por los que se llega a una predicción dada.

Numpy : El principal beneficio de NumPy es que permite una generación y manejo de datos extremadamente rápido. NumPy tiene su propia estructura de datos incorporada llamado arreglo que es similar a la lista normal de Python, pero puede almacenar y operar con datos de manera mucho más eficiente.

G. Seaborn : Es una librería de visualización de datos para Python desarrollada sobre matplotlib . Ofrece una interfaz de alto nivel para la creación de atractivas gráficas.

H. matplotlib : Con esta librería es posible crear trazados, histogramas, diagramas de barra y cualquier tipo de gráfica con ayuda de algunas líneas de código. Se trata de una herramienta muy completa, que permite generar visualizaciones de datos muy detalladas.

Tree : Los árboles de decisión son representaciones gráficas de posibles soluciones a una decisión basadas en ciertas condiciones, es uno de los algoritmos de aprendizaje supervisado más utilizados en machine learning y pueden realizar tareas de clasificación o regresión.

D. Importación de datos

Una de las dificultades que se presenta al realizar el desarrollo de un árbol de decisión es escoger el atributo apropiado, ya que este atributo debe ubicarse como raíz del árbol de decisión para que los demás atributos se generen como descendientes. Este proceso se realiza de forma recursiva en cada nodo. Antes es necesario realizar data reduction y limpieza de datos. Se obtuvieron 10553 datos divididos en 9 columnas. Los datos a importar deben de haber pasado por un proceso de reducción y limpieza, con el fin de detectar registros corruptos, registros imprecisos, datos duplicados o mal formateados.

```
dataset = pd.read_csv("/content/drive/My Drive/IA-2022/ninos_arequipa_V2.csv", index_col=0, encoding='latin-1')
#veamos cuantas dimensiones y registros contiene
dataset.shape
print(dataset)
```

Figura 4. Importación de datos

Posteriormente se verifica la información del dataset.

```
print('Información en el dataset:')
print(dataset.keys())
dataset.info()
```

Figura 5. Verificación del Dataset Preparación de los datos.

```
#variables predictoras
X = dataset.iloc[:,2:7]

#variables a predecir
Y = dataset.iloc[:,0]

#Separo los datos de "train"
#en entrenamiento y prueba para probar los algoritmos
X.head()
```

Figura 6. Preparación de datos

Creación del modelo, se crearon las variables "X" y "Y" para el entrenamiento; las variables "x" y "y" para realizar la prueba.

```
#X_train y Y_train para entrenamiento
# X_test y Y_test para prueba
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.75, random_state=0)
X_train.info()
```

Figura 7. Creación del modelo Validación del entrenamiento.

```
algoritmo = DecisionTreeClassifier(max_depth= 8)
arbol_anemia = algoritmo.fit(X_train,Y_train)

fig =plt.figure(figsize=(25,20))  #dimension
tree.plot_tree(arbol_anemia,feature_names=list(X.columns.values),class_names=list(Y.values), filled=True)
plt.show()
```

Figura 8. Algoritmo para el entrenamiento

Análisis de Resultados

La matriz de confusión nos permite tener un resumen del rendimiento del algoritmo.

```
Matriz_C=confusion_matrix(Y_test, Y_pred)
Matriz_C
```

```
array([[1254,    0,    0,   23],
       [   0,  601,    1,    0],
       [   0,    0,   10,    0],
       [  23,    6,    0, 5997]])
```

Figura 9. Matriz de confusión

```
Precision_G = np.sum(Matriz_C.diagonal())/np.sum(Matriz_C)
Precision_G
```

```
0.9933038534428301
```

Figura 10. Precisión global del modelo

```
Precision_anemia_leve = ((Matriz_C[0,0]))/sum(Matriz_C[0,])
Precision_anemia_leve
```

```
0.9819890368050117
```

Figura 11. Precisión - Anemia Leve

```
Precision_anemia_moderada = ((Matriz_C[1,1]))/sum(Matriz_C[1,])
Precision_anemia_moderada
```

```
0.9983388704318937
```

Figura 12. Precisión - Anemia Moderada

```
Precision_anemia_severa = ((Matriz_C[2,2]))/sum(Matriz_C[2,])
Precision_anemia_severa
```

```
1.0
```

Figura 13. Precisión - Anemia Severa

```
Precision_normal = ((Matriz_C[3,3]))/sum(Matriz_C[3,])
Precision_normal
```

```
0.9951875207434451
```

Figura 14. Precisión - Sin Anemia

Figura 15. Árbol generado

Conclusiones y Trabajos Futuros

Los resultados obtenidos con el modelo de clasificación por árboles de decisión indican que es capaz de generar modelos con los datos que se encuentran en el dataset. Durante este trabajo, existen diversas áreas que quedan abiertas y en las que es posible trabajar. Algunas de estas investigaciones están relacionadas directamente al tema de este trabajo y son resultado de la investigación durante el desarrollo. Otras investigaciones no están relacionadas directamente sin embargo pueden ser retomadas posteriormente o como una opción de exploración a otros investigadores.

A continuación, se presentan algunos trabajos futuros que pueden desarrollarse. Además, se sugieren algunos temas específicos para apoyar y mejorar el trabajo de inteligencia artificial propuesto. Entre los posibles trabajos futuros se destacan:

- Aplicación de Árboles de Decisión en el diagnóstico de Anemia en niños en el Perú.
- Árboles de Decisión en el diagnóstico de Cáncer.

- Calidad de predicción del diagnóstico de anemia.
- Tratamiento de la matriz de confusión para disminuir la tasa de falsos positivos y generar un mejor diagnóstico de la anemia.

Referencias

[1] Jara, Hambre, desnutrición y anemia: una grave situación de salud pública. Revista Gerencia Y Políticas de Salud, 7(15), 7–10. [Online] Available at: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-70272008000200001.

[2] Al-kassab-Córdova, A., Méndez-Guerra, C., & Robles-Valcarcel, P. Factores sociodemográficos y nutricionales asociados a anemia en niños de 1 a 5 años en Perú. Revista Chilena de Nutrición, 47(6), 925–932. https://doi.org/10.4067/s0717-75182020000600925.

[3] O. Munares "Niveles de hemoglobina en gestantes atendidas en establecimientos del Ministerio de Salud del Perú" [online document], 2011. Available: Oxford Reference Online, https://www.scielosp.org/article/rpmesp/2012.v29n3/329-336/es/ [Accessed: Jun 21, 2022].

[4] J. Mendoza, "Sistema experto para alertar y brindar alternativa de tratamiento para la anemia en niños de la provincia de Jaén," M.S. tesis, Facultad de Ingeniería, Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú, 2021.

[5] F. Andrade, "Modelo de regresión Dirichlet bayesiano: aplicación para estimar la prevalencia del nivel de anemia infantil en centros poblados del Perú," M.S. tesis, Escuela de Postgrado, Pontificia Universidad Católica del Perú, Lima, Perú, 2020.

[6] G. Canaza, "Modelo predictivo de riesgo asociado a la anemia en niños menores de 5 años en la Microred Yauri provincia de espinar – Cusco, 2019," M.S. tesis, Facultad de Ingenieria, Estadistica e Informatica, Universidad Nacional del Altiplano, Puno, Perú, 2021.

[7] Sistema de información del Estado Nutricional de niños y gestantes Perú - INS/CENAN (Instituto Nacional de Salud Centro Nacional de Alimentación y Nutrición). Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición. Nov, 2021.

[8] Sistema de información del Estado Nutricional de niños y gestantes Perú - INS/CENAN (Instituto Nacional de Salud Centro Nacional de Alimentación y Nutrición) - Repositorio de Datos - Instituto Nacional de Salud. Ins.gob.pe

[9] Russel, S., & Norvig, P. (2012). Artificial intelligence—a modern approach 3rd Edition. In The Knowledge Engineering Review. https://doi.org/10.1017/S0269888900007724

[10]Dogan, S., & Turkoglu, I. (2008). Iron-deficiency anemia detection from hematology parameters by using decision trees. International Journal of Science & Technology, 3(1), 85-92.

[11]Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. D. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 9(2), 19-24.

[12]Ren, Y., Lu, C., Yang, H., Ma, Q., Barnhart, W. R., Zhou, J., & He, J. (2022). Using machine learning to explore core risk factors associated with the risk of eating disorders among non-clinical young women in China: A decision-tree classification analysis. Journal of Eating Disorders, 10(1). https://doi.org/10.1186/s40337-022-00545-6

[13]Pei, D., Yang, T., & Zhang, C. Estimation of diabetes in a high-risk adult chinese population using j48 decision tree model. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy 13 ,2020.

[14]Bouza, C., & Santiago, A. La minería de datos: árboles de decisión y su aplicación en estudios médicos. Modelación Matemática de Fenómenos del Medio Ambiente y la Salud, 2, 64-78. 2012.

[15]Martínez, G. R. S., & Mejía, J. A. S. Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares. Scientia et technica, 16(49), 104-109, 2011.

[16]Landín Sorí, M., & Romero Sánchez, R. E. Árboles de decisiones para el diagnóstico y tratamiento de pacientes con glaucoma neovascular. Revista Archivo Médico de Camagüey, 16(4), 514-527, 2012.

[17]Zuniga, C., & Abgar, N. Breve aproximación a la técnica de árbol de decisiones. Recuperado de https://niefcz. files.

wordpress. com/2011/07/breve-aproximacion-a-la-tecnica-de-arbol-de-decisiones. pd, 2011.

[18]Tech Target "Dimensionality reduction". Recuperado de https://www.techtarget.com/whatis/definition/dimensionality-reduction

[19]Barla, N. Dimensionality Reduction for Machine Learning. Recuperado de https://neptune.ai/blog/dimensionality-reduction, 2022

Trazabilidad de operaciones en base de datos para mitigar riesgos en los procesos de auditoría

Traceability of database operations to mitigate risks in audit processes

Cesar Mayta Avalos
Universidad Nacional Jorge Basadre Grohmann. Tacna, Perú.
cesar.mayta@unjbg.edu.pe
https://orcid.org/0000-0002-5722-1854

Fernando Rosales Castilla
Universidad Nacional Jorge Basadre Grohmann. Tacna, Perú
fernando.rosales@unjbg.edu.pe
https://orcid.org/0000-0003-0668-2885

Milca Gines Colana
Universidad Nacional Jorge Basadre Grohmann. Tacna, Perú
milca.gines@unjbg.edu.pe
https://orcid.org/0000-0002-3596-2803

ARK: ark:/42411/s9/a58
PURL: 42411/s9/a58

RECIBIDO 24/05/2022 • ACEPTADO 02/07/2022 • PUBLICADO 30/09/2022

RESUMEN

En el ámbito de las bases de datos la falta de la trazabilidad de transacciones u operaciones es vital para responder a incidencias o hechos que pueden originarse dentro de ellas, como la alteración o acceso a información no autorizada. Este artículo busca proponer un modelo de auditoría a fin de mitigar el riesgo, utilizando el enfoque de auditoría de objetos aplicado a tablas y transacciones con Oracle. Finalmente se implementó un laboratorio en el cual se desplegó el modelo propuesto y que permitirá asegurar la confidencialidad, integridad y disponibilidad de la información.

Palabras claves: Auditoría, Bases de datos, Riesgos, Trazabilidad, Oracle.

ABSTRACT

In the field of databases, the lack of traceability of transactions or operations in a database is vital to respond to incidents that may originate within them, such as the alteration of unauthorized information. This article proposes an auditing model to mitigate risk using Oracle's object and

transaction auditing approach. Finally, a laboratory was implemented in which the proposed model was deployed, ensuring the information's confidentiality, integrity, and availability.

Keywords: *Audit, Databases, Risks, Traceability, Oracle.*

INTRODUCCIÓN

Las bases de datos es la columna vertebral de todo sistema de información puesto que si esta fallara todo el sistema sería afectado, por lo que resulta importante que se esté siempre se encuentre disponible, asegurando la continuidad del negocio, por lo que el despliegue de planes y medidas permitirán brindarle la seguridad.

Concretamente en el caso del gestor de base de datos Oracle, vemos que se puede realizar una trazabilidad de operaciones sobre los objetos de esta misma, el cual sirva de control y mecanismo de protección; así durante el análisis y evaluación de riesgos en la aplicación de un proceso de auditoría de sistemas de información garantizara responder a la mitigación de las amenazas que puedan presentarse.

Conceptos previos

Sistema de Gestión de Seguridad de la Información:

El uso de la tecnología ha generado ciertos problemas a las organizaciones, que día tras día son más vulnerables a las amenazas que se presentan en el medio, las cuales pueden llegar a convertirse en un verdadero riesgo para la organización afectando el correcto funcionamiento de las actividades del negocio. Para contrarrestar dichas amenazas, las organizaciones deben generar un plan de acción frente a éstas. Este plan de acción es conocido como Sistema de Gestión de Seguridad de la Información (SGSI) y contiene los lineamientos que deben seguirse en la organización, los responsables y la documentación necesaria para garantizar que el SGSI sea aplicado y genere una retroalimentación.[1]
La definición de SGSI se hace de manera formal en la norma ISO 27001, donde están los estándares y mejores prácticas de seguridad de la información.

Análisis de Riesgos:

El análisis de riesgos es la consideración del daño probable que puede causar en el negocio un fallo en la seguridad de la información, con las consecuencias potenciales de pérdida de confidencialidad, integridad y disponibilidad de la información. En el ámbito de la seguridad

informática, las metodologías de análisis de riesgos conforman una disciplina que se articula desde los Sistemas de Gestión de Seguridad de la Información SGSI en las organizaciones, realizando unos importantes escaneos de vulnerabilidades mediante el uso de una serie de modelos y procesos con el fin de proponer una forma más segura de cuidar la información y los recursos de TI [2].

Algunos de los objetivos de las metodologías de análisis de riesgos corresponden a: Planificación de la reducción de riesgos, prevención de accidentes, visualización y detección de las debilidades existentes en los sistemas y ayuda en la toma de las mejores decisiones en materia de seguridad de la información [3].

Base de Datos:

Una base de datos es una colección de datos relacionados, se construyen siguiendo un diseño y se almacenan datos para realizar acciones específicas. Los datos que se almacenan en una base de datos tienen un origen y pertenecen o llevan relación con un evento en específico de la vida real, asimismo el contenido de las bases de datos es de interés de un grupo de usuarios activos. [4]

Auditoría de base de datos:

La auditoría de bases de datos consiste en un proceso de monitoreo continuo y riguroso de los controles que la administración ha establecido dentro de los sistemas de bases de datos y todos sus componentes para obtener una seguridad razonable de la utilización adecuada de los datos que son almacenados por los usuarios mediante los sistemas de información. El monitoreo y pruebas a los controles determinan la pertinencia y suficiencia de éstos, permitiendo entonces ajustar, eliminar o implementar nuevos controles para asegurar su adecuada utilización.[5]

Modelo propuesto para la Trazabilidad de operaciones mediante la aplicación de auditorías en Oracle

La información constituye uno de los pilares de gran valor de cualquier organización, por lo que resulta importante adoptar mecanismos que permitan asegurar la confidencialidad, seguridad e integridad y que permitan garantizar la continuidad del negocio o servicio.
Es conocimiento que en la actualidad toda organización almacena su información en repositorios y dentro de ellos podemos citar a los sistemas gestores de base de datos ya sean de tipo SQL o noSQL, alojados en una infraestructura de tipo física o en la nube.

Dentro de este contexto, el presente artículo de investigación trata de mostrar de forma descriptiva la importancia que toma el concepto de auditoria de base de datos, enmarcado en

establecer controles que permitan minimizar los riesgos de pérdida de datos, así como poder obtener una seguridad razonable y que permita responder a situaciones de incidencia presentados con la finalidad de explicar algún hecho de revisión y/o investigación.

Las bases de datos como activos de información y de misión crítica, requieren ser protegidas con mecanismos, políticas de seguridad, procedimientos, y controles debidamente verificados y que permitan asegurar la continuidad de los servicios que brinda cualquier organización.
La seguridad de las bases de datos estará garantizada por un modelo de auditoría, configurando el hecho de que, tiene una simbiosis entre estos dos conceptos: "No hay seguridad sin auditoría" [7].

Ubicación del Control:

El modelo propuesto consiste en desplegar el control dentro de la misma base de datos, lo cual será realizado en Oracle aprovechando su componente de auditoría, con la finalidad de establecer y desplegar un modelo auditor que permita asegurar la trazabilidad de las operaciones.

Aplicación de la Auditoría en Oracle:

Con la finalidad de desplegar nuestro modelo de auditoría a fin de garantizar la transparencia en la trazabilidad de las operaciones se realizará en la i) auditoria de objetos, estará centrada específicamente en las tablas del esquema principal de base de datos y que sirve de repositorio de la información; ii) auditoría de transacciones u operaciones de datos, en la cual se desplegará un control mediante disparadores que permita conocer el dato cuando es insertado en la tabla, así como las acciones de alteración y borrado de registros que puedan producirse ya sea desde alguna aplicación de la organización, software propietario de alguna compañía o dentro de la misma base de datos.

Desarrollo del Modelo de Auditoría:

En primera instancia hemos definido que para desarrollar el modelo propuesto de auditoría en Oracle, haremos uso del esquema HR [6], el cual corresponde a un grupo de tablas de ejemplo de Recursos Humanos que viene embebida dentro del sistema gestor de base de datos y adicionalmente viene con información (registros) que nos permitirá describir y aplicar las auditorías.

A continuación presentamos las tablas que este esquema define en la base de datos, así como una breve descripción de cada uno de ellos:

Tabla 1. Tablas del esquema de ejemplo HR

TABLA	DESCRIPCIÓN
REGIONS	Registros que describen la Región para un determinado país
COUNTRIES	Registros que describen un país
LOCATIONS	Registros de Direcciones para una determinada ciudad
DEPARTMENTS	Registros que muestra el Departamento o Área de una Organización
JOBS	Registros de los cargos en una organización así como el sueldo
EMPLOYEES	Registros de empleados asociados a determinado un salario y función
JOB_HISTORY	Registros histórico de una determinado cargo en una organización

a. Activando el componente auditor

La base de datos Oracle, hace uso un componente el cual en primera instancia deberá estar habilitado, esta tarea solo puede ser realizada por un DBA (Database Administrator o Administrador de Base de Datos) por tal motivo, si no se realiza este paso inicial no se podrá desplegar ningún modelo de auditoría que quiera desarrollarse.

```
SQL> show parameter audit_trail

NAME                                 TYPE        VALUE
------------------------------------ ----------- ------------
audit_trail                          string      NONE
```

En caso de encontrarse con el valor NONE, cambiaremos por la activación respectiva y posterior a ello debemos reiniciar los servicios de base de datos, con la finalidad de que el componente de auditoría quede preparado para su uso.

```
SQL> alter system set audit_trail='DB' scope=spfile;
```

b. Descripción del Registro de Auditoría de Objetos

La información que se plasma como auditoría es diversa, por lo que lectura e interpretación de dichos registros responde a datos como: Usuario conectado, identificador de sesión, computadora o servidor desde donde se esta conectando, qué tipo de operación de datos ha realizado como insert – delete – update, y hora/fecha en formato minucioso que trata de explicar dentro de la hora exactamente realizada la acción.

Estos registros de la tabla de auditoría AUDIT TRAIL está basado principalmente en las auditorías propiamente de objetos, por ejemplo: si algún usuario realizó mediante la actualización (update) de un determinado puesto de trabajo de la tabla HR.JOBS y de forma específica en el campo: MAX_SALARY el registro auditado solo registra que usuario realizó tal acción, pero no se podrá visualizar como se encontraba el registro antes de dicha actualización.

A continuación creamos la siguiente política de nombre DML_POL, la cual solo puede ser realizado por un usuario privilegiado:

```
CREATE AUDIT POLICY DML_POL
 Actions
  ALL on HR.DEPARTMENTS,
  ALL on HR.REGIONS,
  ALL on HR.COUNTRIES,
  ALL on HR.LOCATIONS,
  ALL on HR.DEPARTMENTS,
  ALL on HR.JOBS,
  ALL on HR.EMPLOYEES;
```

Por último, dicha política es asignada a los usuarios (usuario _01 y usuario _02) de las aplicaciones los cuales serán monitoreados con auditorías sobre los objetos de tablas de dicho esquema HR:

AUDIT POLICY DML_POL BY usuario _01, usuario _02;

A partir de la ejecución y activación de la política DML_POL el registro en el AUDIT_TRAIL comienza a capturar todas las acciones que puedan efectuar los usuarios monitoreados:

- USUARIO _01: realiza acciones de select y update en la tabla hr.countries
- USUARIO _02: realiza acciones de select sobre la tabla hr.locations

	OS_USERNAME	TERMINAL	DBUSERNAME	CLIENT_PROGRAM_NAME	EVENT_TIMESTAMP	ACTION_NAME	OBJECT_SCHEMA	OBJECT_NAME	UNIFIED_AUDIT_POLICIES
1	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	22/05/22 21:44:33,145000	SELECT	HR	COUNTRIES	DML_POL
2	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	22/05/22 21:44:39,830000	SELECT	HR	COUNTRIES	DML_POL
3	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	22/05/22 21:44:39,877000	UPDATE	HR	COUNTRIES	DML_POL
4	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	22/05/22 21:44:39,893000	SELECT	HR	COUNTRIES	DML_POL
5	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	22/05/22 21:56:56,179000	SELECT	HR	LOCATIONS	DML_POL

Figura 1. Captura del Registro de Auditoría de los objetos de tablas de HR

Como se puede observar en la Figura 1, nos muestra información de auditoría relevante y que responde preguntas:

¿Quién lo realizó? ¿De donde fue realizado? ¿En qué fecha y hora fue ejecutada dicha acción?, ¿Cuales fueron las tablas involucradas? ¿Desde que aplicación fue realizada? ¿Cuál fue la sentencia ejecutada?

c. Descripción del Registro de Auditoría de Operaciones

Este tipo de auditoría hará uso de disparadores (trigger) los cuales serán elaborados por el Administrador de base de datos y seguidamente serán cargados en la base de datos para el uso respectivo.

El objetivo de este tipo de auditoría es poder conocer las acciones que un determinado usuario realiza antes del cambio originado en la tabla principal o crítica, ya sea como una actualización, borrado o alguna inserción, sin embargo tengamos presente que los usuarios hacen uso de aplicaciones por lo las acciones señaladas son desde este origen.

También es importante dar a conocer que puede existir usuarios que tengan acceso en la base de datos por lo que constituye también un hecho importante monitorear sus actividades.

Supongamos que el USUARIO _01 mediante su aplicación esté realizando alguna actualización de un sueldo mínimo (MIN_SALARY) de la tabla HR.JOBS correspondiente al JOB_ID=AD_PRES, por

lo que resulta importante resguarda el valor original antes del cambio, así por ejemplo tenemos lo siguiente:

Row 1	Fields	Info
JOB_ID	AD_PRES	varchar2(10), mandatory, Primary key of jobs table.
JOB_TITLE	President	varchar2(35), mandatory, A not null column that shows job title, e.g. AD_VP,
MIN_SALARY	20080	number(6), optional, Minimum salary for a job title.
MAX_SALARY	40000	number(6), optional, Maximum salary for a job title
ROWID	AAARyDAADAAACR1AAA	

Figura 2. Registro del JOB_ID=AD_PRES antes del cambio MIN_SALARY=20080

Row 1	Fields	Info
JOB_ID	AD_PRES	varchar2(10), mandatory, Primary key of jobs table.
JOB_TITLE	President	varchar2(35), mandatory, A not null column that shows job title, e.g. AD_VP,
MIN_SALARY	21080	number(6), optional, Minimum salary for a job title.
MAX_SALARY	40000	number(6), optional, Maximum salary for a job title
ROWID	AAARyDAADAAACR1AAA	

Figura 3. Registro del JOB_ID=AD_PRES después del cambio MIN_SALARY=21080

Si en este momento, se vuelve a consultar el registro de la tabla HR_JOBS, presentará únicamente el último cambio de 21080 pero no se conocerá el valor inicial antes de este cambio (update) y de solicitarse la trazabilidad de operaciones que pueda haber tenido este registro en la base de datos respecto no podría ser atenderse dicho requerimiento.

Con el despliegue de este tipo de auditoría y luego de la creación del disparador para la tabla HR.JOBS vamos a poder conocer la trazabilidad de operaciones que este registro ha tenido, originado por el USUARIO _01, así al final dicha registro se presenta de la siguiente manera:

Row 1	Fields	Info
SIDU	U	char(1), mandatory
FECHA_SIDU	23/05/22 15:02:52,011000	timestamp(6), mandatory
USUARIO	USUARIO_01	varchar2(20), mandatory
USUARIO_RED	DBSERVER\ADMINISTRADOR	varchar2(50), optional
COMPUTADORA	WORKGROUP\DBSERVER	varchar2(50), optional
IP_TRAZABILIDAD	127.0.0.1	varchar2(20), optional
PROGRAMA	PLSQLDEV.EXE	varchar2(100), optional
JOB_ID	AD_PRES	varchar2(10), mandatory
JOB_TITLE	President	varchar2(35), mandatory
MIN_SALARY	20080	number(6), optional
MAX_SALARY	40000	number(6), optional

Figura 4. Registro auditado de la tabla HR.AUDITANDO_JOBS respecto al JOB_ID=AD_PRES

En el caso desarrollado es importante señalar que se tiene una tabla de HR.AUDITANDO_JOBS la cual tiene la siguiente estructura:

```
CREATE TABLE HR.AUDITANDO_JOBS
( SIDU CHAR(1) NOT NULL,
FECHA_SIDU TIMESTAMP NOT NULL,
USUARIO VARCHAR2(20) NOT NULL,
USUARIO_RED VARCHAR2(50),
COMPUTADORA VARCHAR2(50),
IP TRAZABILIDAD VARCHAR2(20),
PROGRAMA VARCHAR2(100),
---Campos de la tabla HR.JOBS------);
```

Asimismo el disparador (trigger), es el objeto que se encuentra cargado en la base de datos y que permitirá capturar la información de una auditoría, en este caso de la operaciones en la tabla HR.JOBS para efectos de trazabilidad, por lo que se muestra parte del siguiente código y en la cual se ha hecho uso de los verbos en Oracle [8]:

```
CREATE TRIGGER AUDITANDO.DISPARA_HR_JOBS
AFTER INSERT OR DELETE OR UPDATE ON HR.JOBS FOR EACH ROW
BEGIN
INSERT INTO HR.AUDITANDO_JOBS
……..
UPDATE INTO HR.AUDITANDO_JOBS
……..
DELETE INTO HR.AUDITANDO_JOBS
```

Como se puede observar en la Figura 4 y tan igual que la auditoría de objetos, este tipo también nos permitirá responder frente a incidencias o consulta sobre la trazabilidad de un determinado registro

¿Quién lo realizó? ¿De dónde fue realizado? ¿En qué fecha y hora fue ejecutada dicha acción?, ¿Cuáles fueron las tablas involucradas? ¿Desde qué aplicación fue realizada? ¿Cuál fue la sentencia ejecutada?
De alguna forma se apoyará de la auditoría de objetos a fin de establecer y tener mayores elementos de revisión a fin de explicar un determinado suceso, para el caso señalado y que venimos desarrollando podemos visualizar las operaciones que el USUARIO _01 y USUARIO _02 han realizado.

	OS_USERNAME	TERMINAL	DBUSERNAME	CLIENT_PROGRAM_NAME	EVENT_TIMESTAMP	ACTION_NAME	OBJECT_SCHEMA	OBJECT_NAME	UNIFIED_AUDIT_POLICIES
1	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:01:28,947000	SELECT	HR	JOBS	DML_POL
2	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:01:43,920000	SELECT	HR	JOBS	DML_POL
3	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:01:43,967000	UPDATE	HR	JOBS	DML_POL
4	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:01:43,967000	SELECT	HR	JOBS	DML_POL
5	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:02:51,633000	SELECT	HR	JOBS	DML_POL
6	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:02:52,118000	UPDATE	HR	JOBS	DML_POL
7	DBSERVER\Administrador	DBSERVER	USUARIO_01	plsqldev.exe	23/05/22 15:02:52,134000	SELECT	HR	JOBS	DML_POL
8	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	23/05/22 16:32:16,418000	SELECT	HR	JOBS	DML_POL
9	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	23/05/22 16:32:31,647000	SELECT	HR	JOBS	DML_POL
10	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	23/05/22 16:32:42,914000	SELECT	HR	JOBS	DML_POL
11	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	23/05/22 16:32:45,614000	DELETE	HR	JOBS	DML_POL
12	DBSERVER\Administrador	DBSERVER	USUARIO_02	plsqldev.exe	23/05/22 16:33:24,130000	SELECT	HR	EMPLOYEES	DML_POL

Figura 5. Registro auditada a nivel de objeto Tabla

Resultados y discusión

Mediante el modelo de auditoría desplegado en la base de datos en Oracle, se pudo capturar información de las tablas del esquema HR con el cual se desarrolló el laboratorio, consiguiendo el registro de trazabilidad de las operaciones que puede haber realizado un determinado usuario. Estos registros de trazabilidad vienen a conformar una colección de datos auditables, los cuales permitirán minimizar los riesgos relacionados a la alteración y/o eliminación no autorizada de la información.

La evidencia de las auditorías capturadas se logra a partir del modelo desarrollado y desplegado, lo cual demuestra fehacientemente y responde interrogantes como por ejemplo: ¿Quién lo realizó? ¿De dónde fue realizado? ¿En qué fecha y hora fue ejecutada dicha acción?.

Existen programas de auditoría del fabricante Oracle como por ejemplo Oracle Audit Vault and Database Firewall [9] que realizan tareas automatizadas de auditoría, sin embargo el modelo propuesto no solo ofrece información sustancial de auditoría, sino que es una alternativa de menor inversión.

La aplicación de nuestro modelo de auditoría, servirá de cimiento y una estructura sólida para construir un sistema de protección de bases de datos, el cual puede ser consolidado en un único contenedor de datos auditables ya sea en un ambiente de infraestructura local o en la nube [10].

Conclusiones

La propuesta del modelo para implementar una auditoría a nivel de base de datos en Oracle para la trazabilidad de operaciones, viene asociado al hecho de que debemos cautelar y proteger la información en las organizaciones, más aún como elemento importante de la protección de datos personales asociado a los elementos de confidencialidad, integridad y disponibilidad los cuales tiene que ir alineados a disposiciones legales. Por lo que los elementos de seguridad que se implemente tendrá un efecto directo en la calidad de la auditoría de seguridad [11].

En el desarrollo del presente artículo, se ha mostrado de forma intrínseca ciertas amenazas causadas por una configuración predeterminada y si no se adoptan medidas de seguridad para este activo de información, quedará expuesta a elementos internos y externos que pueden causar perjuicio a la organización. El impacto en los datos almacenados lleva consigo el detenimiento de los servicios que se ofrecen, causando daños cuantiosos principalmente en lo económico y de reputación [12].

Referencias

[1] Ladino, Martha Isabel; Villa, Paula Andrea; López, Ana María. Fundamentos de iso 27001 y su aplicación en las empresas. Scientia et technica, 2011, vol. 17, no 47, p. 334-339. [Online]. Disponible en: https://www.redalyc.org/articulo.oa?id=84921327061

[2] M. Doris. Metodologías de la seguridad informática. [On line]. Disponible en: http://seguridadinformatica.bligoo.ec/media/users/22/1142179/files/312461/Metodologia_de_la_Seguridad_Ing.pdf

[3] J. Eterovic y G. Pagliari, Metodología de Análisis de Riesgos Informáticos. [Online]. Disponible en:
http:// www.cyta.com.ar/ta1001/v10n1a3.htm.

[4] Elmasri, R., Díaz Martín, J. M., Navathe, S. B. Fundamentos de sistemas de bases de datos. Madrid: Pearson Educación, 2011.

[5] Murillo, Johnny Villalobos. Auditando en las bases de datos. Uniciencia, 2008, vol. 22, no 1-2, p. 135-140. [Online]. Disponible en: https://www.redalyc.org/articulo.oa?id=475948929017

[6] Modelos y de muestra, "SQL Developer Data Modeler 2.0: scripts DDL de muestra" Oracle, 2022. [Online]. Available:

https://www.oracle.com/cl/database/technologies/appdev/datamodeler-samples.html. [Accessed: May. 22, 2022].

[7] Yang, L. (2009). Teaching database security and auditing. SIGCSE Bulletin Inroads, 41(1), 241–245. https://doi.org/10.1145/1539024.1508954

[8] Database 2 day Developer's, "6 Using Triggers" Oracle, 2022. [Online]. Available: https://docs.oracle.com/database/121/TDDDG/tdddg_triggers.htm#TDDDG50000 [Accessed: May. 23, 2022].

[9] Oracle. (2017). Oracle Audit Vault and Database Firewall. March. http://www.oracle.com/technetwork/database/database-technologies/audit-vault-and-database-firewall/overview/index.html

[10]O. Cinar, RH Guncer y A. Yazici, "Seguridad de bases de datos en nubes privadas de bases de datos", Conferencia internacional sobre ciencia y seguridad de la información (ICISS) de 2016, 2016, págs. 1 a 5, doi: 10.1109/ICISSEC.2016.7885847.

[11] -ul-Hasan, M., & Othman, S. H. (2019). A Conceptual Framework of Information Security Database Audit and Assessment. International Journal of Innovative Computing, 9(1), 7–13. https://doi.org/10.11113/ijic.v9n1.206

[12]García, M. J. (2013). Database Main Threats Analisys Using MS SQL Server. 1–5. http://www.unab.edu.co/sites/default/files/MemoriasGrabadas/papers/capitulo9_paper_10.pdf

Clasificación de tutoriales en YouTube basándonos en el análisis de sentimientos realizados a sus comentarios

Classification of tutorials on YouTube based on the analysis of feelings made to your comments

Valeria Alejandra Goyzueta Torres
Universidad La Salle. Arequipa, Perú.
agoyzueta@ulasalle.edu.pe

Ronald Fabricio Centeno Cardenas
Universidad La Salle. Arequipa, Perú.
rcentenoc@ulasalle.edu.pe
https://orcid.org/0000-0001-6639-8603

Victor Andre Ranilla Coaguila
Universidad La Salle. Arequipa, Perú.
vranillac@ulasalle.edu.pe

ARK: ark:/42411/s9/a66
PURL: 42411/s9/a66

RECIBIDO 10/07/2022 • ACEPTADO 22/08/2022 • PUBLICADO 30/09/2022

RESUMEN

El flujo de información surge día a día mediante internet de manera continua gracias a las constantes interacciones presentes entre los usuarios, estas interacciones presentan sentimientos que pueden ser positivos o negativos. Esto ayuda mucho a los creadores de contenido de las redes sociales a comprender cuan útil es lo que ellos hacen para sus seguidores, y es que, si estos son un gran número, un análisis hecho por una sola persona no es suficiente. Para ello es necesario el uso de herramientas que operan con grandes cantidades de datos como BERT, que es un modelo que ayuda al análisis de sentimientos y clasificación de comentarios basados en lo que expresa uno de estos. En este trabajo se usará este modelo para la clasificación de comentarios de YouTube y clasificación de videos de esta misma plataforma, valorando estos videos según su contenido y ayudando a los espectadores a elegir los videos si es que estos lo ayudarán con respecto a lo que se encuentran buscando. Se harán además uso de métricas y de sugerencias futuras para la propuesta mencionada en este trabajo.

Palabras claves: Análisis de sentimientos, Comentarios de Youtube, Clasificación de videos.

ABSTRACT

The flow of information arises day by day through the Internet in a continuous way thanks to the constant interactions between users, these interactions present feelings that can be positive or negative. This helps social media content creators a lot to understand how useful what they do is for their followers, and if these are a large number, an analysis done by a single person is not enough. For this, it is necessary to use tools that operate with large amounts of data, such as BERT, which is a model that helps analyze sentiments and classify comments based on what one of them expresses. In this work, this model will be used for the classification of YouTube comments and the classification of videos on this same platform, evaluating these videos according to their content and helping viewers to choose the videos if they help them concerning what is expected. find searching. This work will also use future metrics and suggestions for the proposal.

Keywords: *Sentiment Analysis, Youtube Comments, Video Ranking.*

INTRODUCCIÓN

La capacidad de adaptación de la tecnología sobre diferentes contextos, no deja atrás a ámbitos como la enseñanza y el aprendizaje, pues permite tanto a estudiantes como a maestros ser partícipes del intercambio de conocimiento sin importar la distancia ni las limitaciones físicas presentes. Este conocimiento viaja de plataforma en plataforma y está siempre presente en donde más interacción entre personas ocurre, siendo en la actualidad las redes sociales.

La vida en las redes sociales es muy amplia y compleja de entender, pero puede resumirse como la interacción entre diversos usuarios que intercambian información, intereses y opiniones sobre un tema en específico de manera remota. Esta interacción se lleva en mayor medida dentro de los comentarios de las publicaciones que realizan otros usuarios, llevando a dar una visión general sobre una opinión de un tema en específico que muchas veces no es muy acertada, pues la cantidad de información dentro de este mismo es muy diversa.

Dentro de las plataformas donde más se realiza este intercambio de opiniones se encuentran: Facebook, Twitter y YouTube, siendo esta última considerada la plataforma más grande de videos en la red, donde se realizan subidas e intercambio de contenido, tanto educativo como de entretenimiento. En ella cada minuto se sube un aproximado de 500 horas de videos y más de un billón de videos dentro de la misma son visualizados en diversos lugares del globo [2], dentro de los cuales se encuentran los videos tutoriales.

Los videotutoriales son una herramienta que ayudan al fortalecimiento de conocimiento, como la aclaración de dudas sobre un tema en específico, que puede ser adquirido en un entorno

presencial o simplemente sea un nuevo concepto que quiere ser aprendido. Esta clase de material tiene un mismo formato: Un tutor enseñando acerca de un tema que domina, haciendo uso de herramientas para facilitar la comprensión de sus instrucciones y convirtiendo su video en un material reconocido y adecuado de enseñanza y aprendizaje.

Gracias al alcance de internet, la propagación de estos materiales se realiza rápidamente y más en YouTube, pues es mucho más sencillo transmitir conocimientos mediante videos que haciendo empleo de otros sentidos. Cada video que es subido a YouTube es categorizado basándose en su contenido, y calificado según ese mismo sobre la base de los usuarios que interactúan con él.

Todo esto gracias a tres herramientas brindadas por la misma plataforma: los botones de like, dislike y los comentarios. La calidad de los videos es determinada por estos indicadores y marcan una reputación sobre quien subió este material. Las principales fuentes que determinaban esta calidad eran las cantidades que eran brindadas por los botones previamente mencionados, pues estos determinaban los niveles de aprobación que poseía un video y cuán útil es con respecto al ámbito donde este se enfocaba. Esto hasta el año 2021 se deshabilitó la visualización de la cantidad de estos botones, dejando a los usuarios con una vaga idea de la utilidad de un video a simple vista.

Pero haciendo un análisis más profundo, los comentarios dentro de los videos son los que también determinan la reputación de cada creador y la calidad de cada video, pues cada comentario contiene información valiosa que puede ayudar a la clasificación de un video y su relevancia en la plataforma.

Cada comentario contiene palabras clave que ayudan a identificar una emoción asociada a una respuesta de reacción al video, que puede ser tanto positiva como negativa. El hecho de identificar estas palabras asociadas a emociones en sencillo, pero cuando la cantidad de datos es exponencial, pues estamos hablando de plataformas globales, es necesario el uso de algoritmos y métodos de procesamiento de lenguaje natural. Al realizar la segmentación de comentarios en dos categorías, basados en palabras clave relacionadas con sentimientos de usuarios, indica la relación directa con un área del procesamiento del lenguaje natural (NLP) llamada Análisis de Sentimientos (SA). Como su nombre lo indica, su principal objetivo es la extracción de sentimientos dentro de comentarios, para esto los comentarios deben ser clasificados con base en información puntual u opiniones subjetivas [5].

El análisis de sentimientos normalmente está compuesto de 4 fases: Extracción de información, Procesamiento de data, Clasificación de sentimientos y Presentación de la salida, dentro de los cuales el Procesamiento de data es el paso que más esfuerzo requiere, pues envuelve procesos como: Preprocesamiento de texto, Feature Extraction y Feature Selection. Una vez completada esta fase, la clasificación hace uso de algoritmos de Machine Learning (ML) para realizar una clasificación de polaridad de comentarios, que ayudará a la clasificación de videos según su

utilidad. Este estudio se enfoca en la obtención de opiniones basadas en los comentarios que son manifestados a manera de reacción a los videos tutoriales que son subidos a la plataforma de YouTube.

Para poder determinar su utilidad y realizar la clasificación según esta misma (útil, inútil), además de representar las clasificaciones de comentarios en variables numéricas. Para ello, primero se realizará un preprocesamiento de información, dejando aquellos comentarios que presenten sentimientos positivos o negativos explícitamente involucrados luego haciendo uso de algoritmos de procesamiento de lenguaje natural, determinar el porcentaje predominante y concretar la primera tarea enunciada, la tarea de clasificación.

Este trabajo está organizado de la siguiente manera: La introducción, motivación del trabajo, los trabajos previos que han sido ejecutados en el área del procesamiento del lenguaje natural, el marco teórico, la propuesta que involucra parte de la implementación, los resultados y finalmente las conclusiones y recomendaciones aplicables en trabajos posteriores relacionados a este sector computacional.

II. Motivación

El siguiente trabajo busca desarrollar una herramienta de análisis de sentimientos de comentarios que surgen como respuesta a videos tutoriales que son subidos a la plataforma de YouTube, para enseñar las palabras más predominantes que indiquen la utilidad de un video a base de su contenido y la opinión popular que es generada por los usuarios. Las preguntas que se intentan resolver, mediante la elaboración de este trabajo, son las siguientes:

- P1: ¿Qué palabras son las más frecuentes cuando un video es considerado útil por los usuarios?
- P2: ¿Qué palabras son las más frecuentes cuando un video es considerado inútil por los usuarios?
- P3: ¿Qué porcentaje de comentarios presentan una posición ’neutral’ o ’indefinida’?

III. Trabajos relacionados

Para la elaboración de este trabajo hemos recolectado trabajos previos, cuya intención se centra en el análisis de sentimientos con base en comentarios emitidos como reacción a videos publicados en YouTube: Hanif et al. [3] elaboran un modelo basado en NLP, que se encarga de retornar a los usuarios los videos más relevantes y populares dependiendo de los comentarios de las personas.

Esta propuesta hace uso de herramientas de preprocesamiento de texto encargadas de extraer aquellos caracteres y conjuntos de palabras que son totalmente irrelevantes, como enlaces, símbolos, caracteres, emoticones y aquellos comentarios que no se encuentran en el idioma al que está orientada la propuesta para poder eliminarlos, además de los signos de puntuación.

Una vez eliminados los caracteres que no sirven para la propuesta, se procede a generar un dataset limpio, con la forma singular de las palabras que conforman los comentarios. A todos los adjetivos que conforman los comentarios se les aplicó un POS Tagger, para generar un segundo dataset. Sobre la base de estos dos datasets generados, se realizó el análisis de sentimientos, que presenta como limitación o como condición determinante del rendimiento, la forma en la que los comentarios y las palabras que los conforman son procesadas y el análisis semántico de su contenido.

Otra de las propuestas que hacen análisis de sentimientos dentro de comentarios hacia material audiovisual es la propuesta elaborada por Obadimu et al. [11] que si bien no hacen referencia a si los comentarios son positivos o negativos como tal, hace énfasis en el análisis de las palabras que representan a un sector, que en este caso es la toxicidad dentro de los comentarios de YouTube sobre la opinión de una corriente política. Ellos toman en consideración 5 tipos de toxicidad.

Haciendo uso de una CNN[1], se logró determinar cuándo un comentario era tóxico dentro de una discusión entre usuarios. Para tokenizar las palabras que forman parte de los comentarios se hizo uso de la librería de Python: NLTK. Que es ampliamente utilizada en el área del procesamiento del lenguaje natural. El trabajo elaborado por Obadimu et al. presenta sugerencias de sanción frente a esta clase de comentarios identificados dentro de Youtube. Cunha et al. [1] nos presenta una manera de clasificar los comentarios de un video con base en la influencia que tienen sobre los usuarios, la relevancia del video y la calidad visual de este mismo.

Cada clasificación presenta tres posibles posiciones: positivo, negativo y neutral. Esta propuesta hace uso de heurísticas de preprocesamiento de texto para luego aplicar Deep Learning en la predicción de las reacciones de los usuarios a ciertos videos manifestadas dentro de los comentarios. Los autores sugieren que un preprocesamiento más exhaustivo puede ser necesario para poder mejorar la efectividad del modelo propuesto. Singh y Tiwari [12] nos presentan una forma de realizar el análisis de comentarios de Youtube haciendo uso de diferentes técnicas de Machine Learning y Deep Learning. Además de librerías de Python como: SciKitLearn que ayuda a la conversión de la data textual a numérica para poder interactuar con ella a manera de vectores. Ellos hicieron uso de tareas clásicas del procesamiento del lenguaje natural: Lemmatisation y la remoción de caracteres que no aportarían nada al análisis de sentimientos, como los signos de puntuación. Para la parte de clasificación, se usaron diversos algoritmos, entre

[1] CNN : Convolutional Neural Network (Red neuronal convolucional).

ellos los 6 más conocidos: Bayes Naives, Support Vector Machine, etc. Haciendo una comparación entre ellos, se llegó a la conclusión de que el algoritmo que arroja mejores resultados es el de Random Forest, y el que resultados más bajos logró obtener fue el de Naive Bayes, que necesitó menos preprocesamiento de texto. Muhammad et al. [10] presenta una clasificación a los comentarios de los videos de YouTube combinando los métodos de Naïve Bayes y Support Vector Machine (NBSVM) con un enfoque de Clasificación Binaria.

El uso de estos métodos fue elegido por ellos porque Naïve Bayes es muy bueno en la clasificación de textos con un pequeño número de datos, mientras que Support Vector es muy bueno en la clasificación de textos con un número relativamente alto de datos. Los resultados obtenidos muestran que la combinación de Naïve Bayes y Support Vector Machine produce un mejor nivel de precisión y un mayor rendimiento. Sin embargo, la combinación de varios clasificadores no siempre aumenta la precisión de las clasificaciones.

IV. Marco teórico

Inteligencia Artificial

La inteligencia artificial hace referencia a sistemas informáticos con la capacidad de hacer predicciones o realizar acciones basándose en los patrones de los datos disponibles y poder aprender de sus errores para ser más precisos [13]. Una inteligencia artificial avanzada procesa la información nueva con suma rapidez y precisión, es por ello que generalmente se puede asociar el entendimiento humano a la computadora por medio de este tipo de tecnologías.

Procesamiento de lenguaje Natural[2]

El procesamiento del lenguaje natural es un enfoque de la inteligencia artificial que ayuda a interpretar el lenguaje humano a través de algoritmos de análisis de texto y reconocimiento de texto [13], haciendo uso de elementos de ciencia y lingüística computacional, para que el lenguaje humano sea procesado bajo una correcta comprensión por parte del computador.

Su importancia radica en ayudar al entendimiento entre un computador y una persona a través del lenguaje humano, cooperando así en la realización de múltiples tareas basadas en reconocimiento de voz, interpretación y análisis, además de la medición del sentimiento.

[2] NLP : Natural Language Processing (Procesamiento natural del lenguaje).

Medición de sentimiento

El análisis de sentimiento hace referencia al uso de NLP, por medio de herramientas basadas en lingüística computacional y análisis de texto, para reconocer y sustraer información relacionada con los recursos analizados. La medición se realiza basándonos en un tratamiento enfocado en relaciones estadísticas y de asociación lingüística que repara en la creación de conclusiones referidas a encontrar una meta u objetivos[4].

Para realizar un análisis o medición de sentimiento se utilizan Datasets[3] basados en reseñas, opiniones y comentarios, que brindaran una idea determinada al enfoque de análisis impuesto.

Datasets

El término DATASET se refiere a un archivo que contiene uno o más registros de información. Estos registros son seleccionados y clasificados a base de al enfoque de su aporte. Muchas veces estos registros de información se utilizan para almacenar información que necesitan las aplicaciones o el propio sistema operativo; Al final se catalogan basándonos en el tipo de información al cual se enfocan.

Al término de su preprocesamiento y su clasificación, son procesados con métodos y técnicas de inteligencia artificial basados en algoritmos CNN, BOW[4] o BERT[5] , para analizar y concluir la correcta interpretación de los datos procesados.

Convolutional Neural Network

Son un tipo de redes neuronales artificiales donde las neuronas corresponden a campos receptivos de una manera muy similar a las neuronas en la corteza visual primaria de un cerebro biológico. Este tipo de red es una variación de un perceptrón multicapa, sin embargo, debido a que su aplicación es realizada en matrices bidimensionales, son muy efectivas para tareas de visión artificial, como en la clasificación y segmentación de texto e imágenes, entre otras aplicaciones [13].

Las CNN aprenden por lo general a reconocer una diversidad de objetos dentro de imágenes [13], además de interpretar y entender textos por medio de oraciones y palabras específicas, pero para ello necesitan entrenarse con una cantidad importante de muestras, las cuales pueden ser

[3] Dataset : Conjunto de datos.
[4] BOW : Bag of Words.
[5] BERT : Bidirectional Encoder Representations For Transformers.

muchas veces brindadas por un gran banco de información, generalmente conocidos como Big data.

Big data

Big data se refiere a conjuntos de datos que son demasiado grandes o complejos para ser tratados por el software de aplicación de procesamiento de datos tradicional. Los desafíos del análisis de big data incluyen la captura de datos, el almacenamiento de datos, el análisis de datos, la búsqueda, el intercambio, la transferencia, la visualización, la consulta, la actualización, la privacidad de la información y la fuente de datos [13].

El uso actual del término big data tiende a referirse al uso de análisis predictivos, análisis del comportamiento del usuario u otros métodos avanzados de análisis de datos que extraen valor de los grandes datos, y rara vez a un tamaño particular de Datasets[4].

Herramientas más populares en el uso de análisis de sentimientos

El análisis de sentimientos por lo general consiste en valorar y estimar la disposición de un usuario, en relación con sus opiniones; con la finalidad de obtener información que permita comprender su postura y reacción respecto a un servicio o producto en específico.

Por ello se utilizan diversas herramientas de análisis de sentimiento relacionadas con tecnologías avanzadas de inteligencia artificial, entre las que se encuentran el enfoque de procesamiento del lenguaje natural, análisis de frases o textos y data science[6] [7], pero entre las que destacan son el uso de estos enfoques a través de redes neuronales recurrentes.

Es por ello que la mayoría de estas herramientas son desarrolladas de forma modular y su entrenamiento es desarrollado por la comunidad open source[7], obteniendo así diversos métodos de trabajo basados en el análisis del lenguaje natural para la extracción de datos y el desarrollo analítico de sintaxis basándonos en entidades, la detección de sentimiento y la clasificación de contenido.

Entre las herramientas más reconocidas para el desarrollo de NLP con redes neuronales recurrentes encontramos:

[6] Data science : Ciencia de datos para identificar, extraer y estudiar información subjetiva.
[7] Open Source : Software de código abierto y desarrollo libre.

a) Gooogle CLoud Platform - Natural Language IA[8]

Como su nombre lo refiere, Google creo una herramienta para obtener información de textos no estructurados mediante el aprendizaje automático del algoritmo de Google; todo ellos con el objetivo de realizar un análisis de texto perspicaz con el aprendizaje automático que extrae, analiza y almacena texto, entrene modelos personalizados de aprendizaje automático sin una sola línea de código con AutoML y aplique la comprensión de lenguaje natural a las aplicaciones con API de lenguaje natural.

En síntesis, esta herramienta utiliza el análisis de entidades para encontrar y etiquetar campos dentro de un documento, luego las analiza para comprender las opiniones de los usuarios, encontrando información procesable sobre servicios y productos.

b) Open IA GPT3

Open IA desarrollo una ampliación de los modelos de análisis de lenguaje natural que mejora el rendimiento de desarrollo de tareas y pocos intentos. Desarrollo GPT3, un modelo de lenguaje autorregresivo con millones de parámetros que ayuda a determinar una respuesta sugerente a la determinación de sentimientos en el desarrollo de análisis de opiniones y predicciones[6].

c) BERT

Bert significa, representaciones de codificador bidireccional de Transformers y es un modelo de aprendizaje automático utilizado para tareas NLP. Fue entrenado con Wikipedia en inglés y BookCorpus; actualmente existen dos variaciones de BERT preentrenadas, el modelo base de 12 capas neuronales y otro de 24. En rendimiento, BERT es realmente superado por GPT3 de Open IA, pero el acceso limitado a GPT3 obliga a utilizar el enfoque BERT.

V. Metodología

Propuesta

Para este trabajo, un dataset de comentarios de videos de YouTube que son catalogados y juzgados por su contenido ha sido utilizado [9]. Para continuar con las tareas propuestas, un proceso de limpieza ha sido necesario para poder obtener la información en un formato más legible y entendible, además de ordenado al momento de realizar la tarea de clasificación.

[8] Google Cloud Platform : https://cloud.google.com/natural-language

La tarea de limpieza se llevará a cabo con librerías de procesamiento de lenguaje natural y Machine Learning, eliminando aquellos caracteres que no son de utilidad e interrumpe el entendimiento del comentario.

Para luego obtener los comentarios filtrados y realizar un análisis de sentimientos que permitirá la clasificación de comentarios en alguna de las tres clases: positivo, neutral o negativo. Y con base en esta clasificación poder clasificar un video según la cantidad de registros presentes en una clase Figura 1.

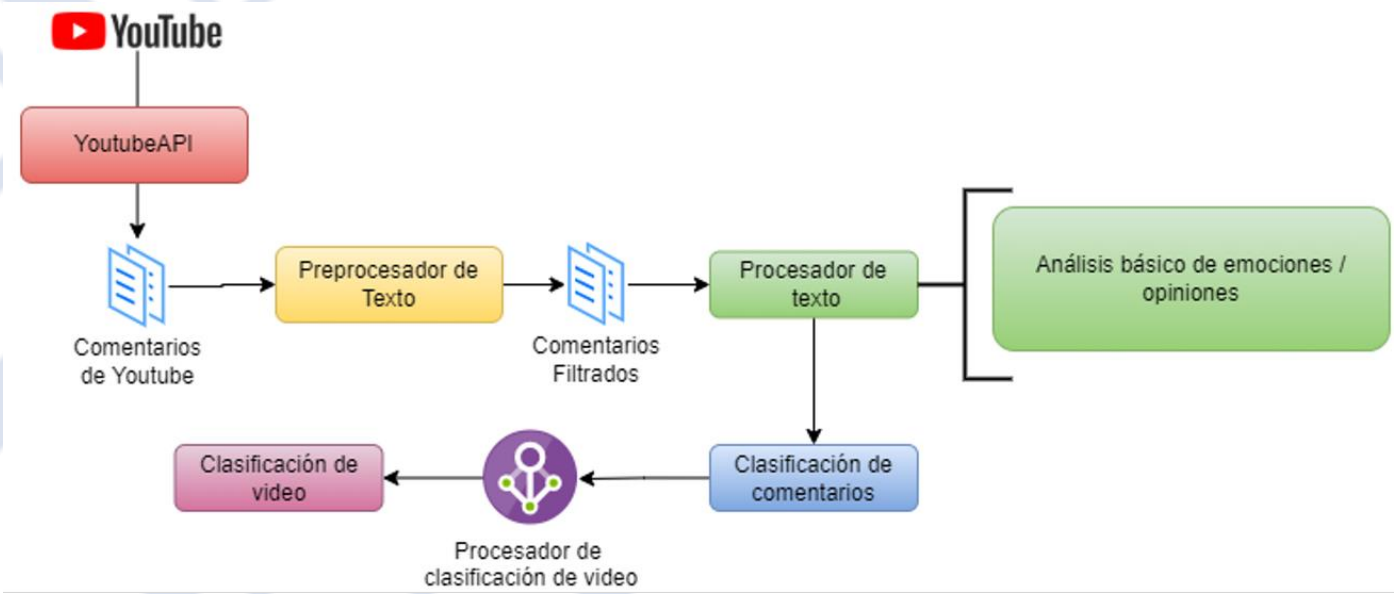

Figura 1, Pipeline de la propuesta.

Descripción de la data

Hemos usado este dataset que se encuentra disponible desde el año 2020, y cuya última actualización se llevó a cabo ese mismo año. Estos comentarios han sido extraídos de la Play Store de Google, y se enfocan en comentarios de crítica a aplicaciones, que en general suelen ser más de 82 billones de aplicaciones. La Tabla 1, enseña los atributos de dataset.

Tabla 1. Descripción de atributos

Atributos	Descripción
reviewID	ID del comentario
UserName	Usuario
UserImage	Perfil de Usuario
Content	Contenido de la aplicación
thumbsUpCount	'likes' hacia el comentario
at	La aplicación
replyContent	Respuestas de un comentario
repliedAt	Comentario respondido

El archivo pesa un aproximado de 3.45Mb, conteniendo más de 12000 registros dentro de el, muestra además una clasificación basa en números del 1 al 5, representado en la Figura 2.

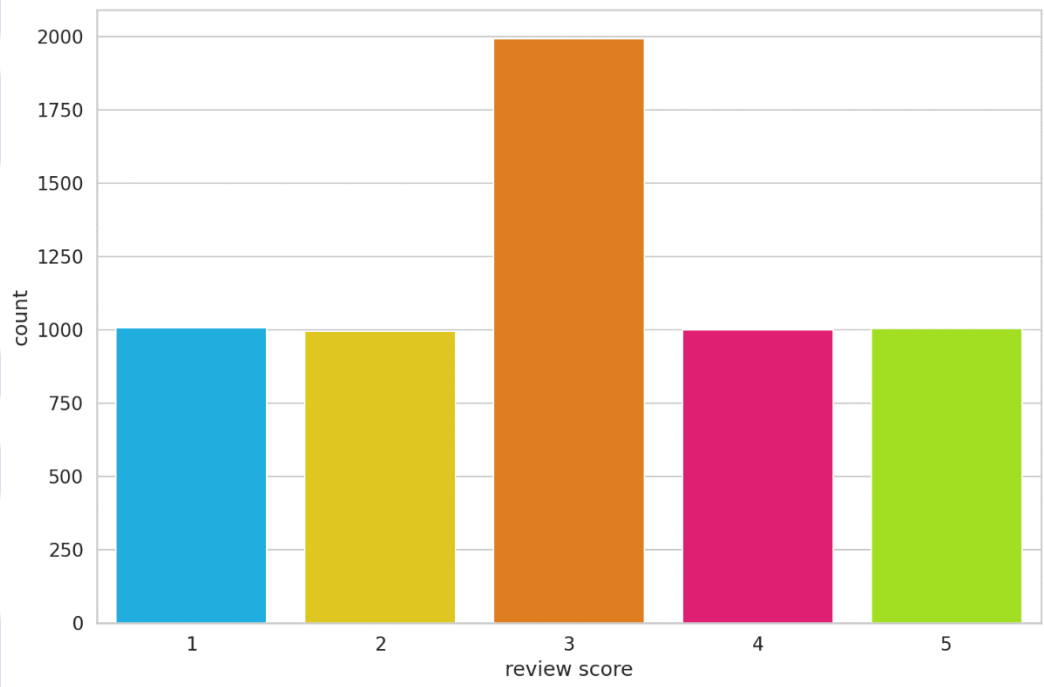

Figura 2, Clasificación numérica de los comentarios acerca de una aplicación.

Herramientas

Para la implementación de la propuesta mencionada anteriormente se ha decidido hacer uso de las siguientes herramientas:

1. BERT: Ya mencionado anteriormente, la funcionalidad de este modelo bidireccional de aprendizaje es de gran utilidad, pues entre las principales tareas que realiza se encuentra la de

clasificación de texto basada en los sentimientos que se encuentran dentro de los comentarios. Con esta herramienta, dentro de la implementación de esta propuesta se ha usado de manera que podamos clasificar los comentarios en tres clases, obteniendo la clase predominante y clasificando de esta manera el videotutorial que nosotros hemos seleccionado. Siendo mucho más específicos la manera en la que BERT es usado en la implementación de la clasificación de comentarios y, por tanto, de videos es en la división de comentarios con base en su contenido que ayuda a la clasificación de videos como tales.

2. API de YouTube: Esta herramienta es de gran utilidad, pues ayuda a conseguir los comentarios, de manera que haciendo uso del resto de herramientas podemos clasificarlos. Podemos decir que esta herramienta es la fuente principal de datos que ayudan a poner a prueba lo que se ha implementado, haciendo que la propuesta mencionada dentro de este trabajo sea aplicable dentro de entornos en la vida real.

3. Python: Lenguaje de programación interpretado de tipado fuerte y dinámico, que soporta la programación orientada a objetos, la programación funcional y la programación imperativa, usado en áreas como el Machine Learning, Deep Learning, Reconocimiento Facial, y otras tareas más. Se usa para la construcción de variedad de aplicaciones usadas en cualquier contexto.

Librerías

- **Numpy:** Librería de Python que ayuda en la creación de vectores y matrices de grandes proporciones, acompañada de una gran cantidad de funciones de alto nivel que ayudan en la interacción con estas estructuras numéricas.
- **Pandas:** Biblioteca usada para el análisis y manipulación de datos. Es una extensión de Numpy. Ayuda en la definición de nuevas estructuras de datos que pueden ser accedidas mediante índices o lo que pueden ser nombres para las columnas y ubicaciones para las filas. Permite además la operación con estos datos de manera rápida y eficiente.
- **Matplotlib:** Es una librería para la generación de gráficas que surgen a base de arrays o listas definidas dentro del lenguaje de programación Python. Contiene variedad de gráficas y ayudan a la expresión de métricas de manera visual, funcionando de esta manera como un complemento para el resto de librerías cuyo centro principal es la manipulación directa con información en grandes volúmenes.
- **Seaborn:** Similar a Matplotlib, es una librería que se basa en la ayuda de representación visual de información en gráficas, el gráfico más usado al utilizar esta librería es el histograma.
- **Torch:** Es una librería open-source enfocada en el Deep Learning y aprendizaje automático que acelera el camino desde la creación de prototipos de investigación hasta el despliegue de las aplicaciones en un entorno de producción.

- **Transformers:** Proporciona APIs para descargar y entrenar fácilmente modelos preentrenados de última generación. El uso de estos puede ayudar en la reducción de costo computacional, la huella de carbono y ahorrarle tiempo en lugar de realizar un entrenamiento desde cero de un modelo que no ha sido entrenado con anterioridad.
- **SciKitLearn:** Es una librería de aprendizaje automático que soporta algoritmos de clasificación, regresión y clustering: SupportVector Machines, Random Forests, Gradient Boosting, K-means y DBSCAN). Está diseñada para el poder interactuar con las librerías Numpy y Scipy en el ámbito del aprendizaje profundo y automatizado.

VI. Resultados y Comparativa

Al hacer uso de dataset ya mencionado, la división que se realiza entre la data para el entrenamiento y la data que va a ser evaluada basándonos en el entrenamiento se encuentra en una proporción del 80 % dejando el resto de data para la evaluación de la efectividad del modelo. Para poder verificar la calidad y el éxito de este modelo se utilizarán métricas que permiten evaluar la efectividad de la propuesta de este trabajo:

- **F1-Score:** Es la medida de precisión que tiene un test, hace uso de métricas previas como lo son la precisión y la exhaustividad del modelo. Opera con los falsos negativos, falsos positivos, verdaderos positivos y verdaderos negativos. Suele ser empleado en la fase de prueba de algunos algoritmos, sobre todo en aquellos de clasificación (Ecuación 1)

$$F1 - Score = \frac{2 * presicion * recall}{presicion + recall} \quad (1)$$

- **Accuracy:** O precisión, mide la cantidad de predicciones correctas en relación con el total de predicciones realizadas (Ecuación 2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- **Matriz de confusión:** Indica la cantidad de elementos que han sido clasificados en sus respectivas clases en forma de matriz. De allí su nombre.

Se ha definido el espacio de trabajo con un aproximado de 6000 registros, que tienen comentarios variados, pertenecientes a las tres clases que se han mencionado antes. Como cada comentario posee un puntaje que puede estar entre los números 1 a 5, se ha colocado estos indicadores

numéricos en indicadores textuales, para poder hacer la tarea de clasificación más sencilla y basada enteramente en clases que puedan ser entendidas y no ambiguas:

- 1-2: Negativos
- 3: Neutrales
- 4-5: Positivos

La distribución de registros se visualiza en la Figura 3.

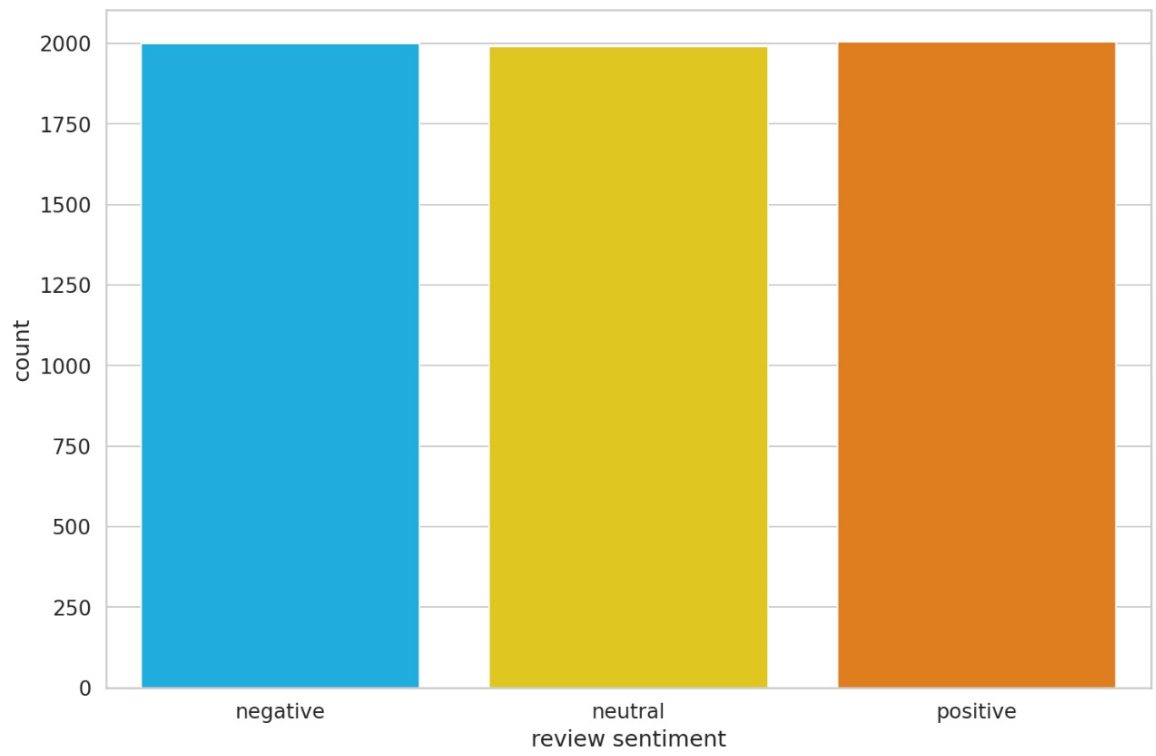

Figura 3, Distribución de registros en tres clases textuales.

Una vez establecidas las tres clases, se procede a hacer la medición del accuracy del modelo, cuyo valor numérico es de 0.73 o representado como el 73 %. Y finalmente se procede a calcular el F1-score que como se ha mencionado hace uso del accuracy y el recall, donde el puntaje es: 0.72 o representado como el 72 %. De la misma manera, la matriz de confusión presenta un alto índice de predicción correcta en aquellos comentarios neutrales, como se refleja en la Figura 4.

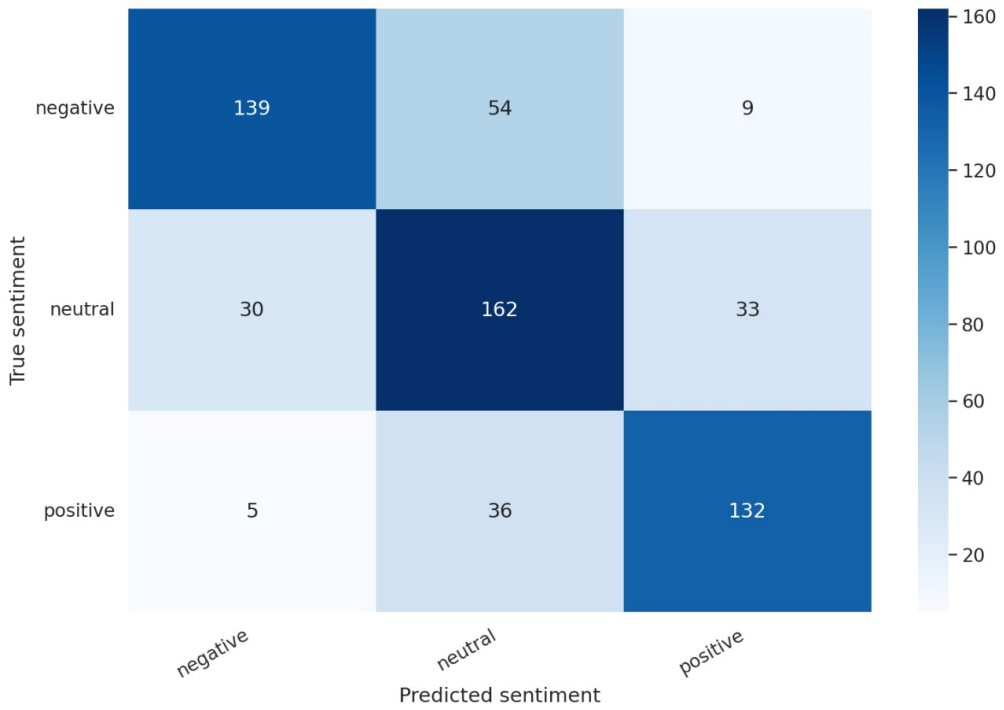

Figura 4, Matriz de confusión del modelo

Los datos obtenidos han sido comparados con [8], el cual nos muestra una tabla de resultados obtenidos usando dos métodos de extracción de caracteres, Term Frequency-Inverse Document Frequency (TFIDF) y vectores de documentos (Doc2vec), y usando métodos de clasificación como Random Random Forest (RF), Bootstrap Aggregating (Bagging), Gradient Boosting (GBT), Naïve Bayes (NB), Ridge Regression y Linear Support Vector Machine (SVC). Estos datos han sido comparados con nuestro resultado obtenido y la comparación se muestra en la Tabla 2.

Tabla 2. Resultados de diferentes tipos de clasificación.

Algoritmo		Accuracy
TF-IDF	RF	71.49 %
	Bagging	71.57 %
	GBT	77.19 %
	NB	65.53 %
	Ridge	60.29 %
	SVC	64.61 %
Doc2Vec	RF	63.83 %
	Bagging	64.07 %
	GBT	69.23 %
	NB	59.58 %
	Ridge	69.84 %
	SVC	70.11 %
BERT		72 %

Aplicación en los comentarios de Youtube

Una vez entrenado en modelo, procederemos a aplicar este mismo en la clasificación de comentarios, para la posterior clasificación de los videos, basándose en la cantidad de comentarios predominantes que posea un video. La predominancia de los comentarios ayudará a clasificar un video como útil, parcialmente útil y no útil, pues hablamos de tutoriales y estos deben de poseer utilidad que será aplicada por el resto de usuarios que miren el video.

Para la obtención de los comentarios se ha usado una API de YouTube como ya se ha mencionado, esto nos ayudará gracias a la implementación que hemos realizado a que obtengamos específicamente solo los comentarios y analicemos la cantidad de sentimientos que estos poseen de trasfondo. Para esta prueba se han usado tres videos relacionados a la computación y que contienen variedad de comentarios que se encuentran dentro de las clases que hemos establecido con anterioridad.

Los tres videos que vamos a usar son los siguientes:

- Video 1: Screensaver Not Working in Windows 10 FIX [Tutorial]
- Video 2: How to Enable keyboard in BIOS. 100 % working (HD)
- Video 3: Windows 11 Blue Screen Error Critical Process Died FIX [Complete Solution]
- Video 4: React Hooks Course - All React Hooks Explained

Cuyos resultados se encuentran reflejados en la Tabla 3, 4, 5 y 6.

Tabla 3. Resultados del Video 1

Tipo de comentario	Cantidad
Positivo	14
Negativo	16
Neutral	19

Tabla 4. Resultados del Video 2

Tipo de comentario	Cantidad
Positivo	10
Negativo	5
Neutral	7

Tabla 5. Resultados del Video 3

Tipo de comentario	Cantidad
Positivo	33
Negativo	30
Neutral	82

Tabla 6. Resultados del Video 4

Tipo de comentario	Cantidad
Positivo	267
Negativo	39
Neutral	195

Dejando a los videos con la siguiente clasificación:

- Video 1: Parcialmente útil.
- Video 2: Parcialmente útil.
- Video 3: Parcialmente útil.
- Video 4: Útil.

Conclusiones

Gracias al uso de herramientas que interactúan con el procesamiento del lenguaje natural como lo es BERT se pueden obtener los sentimientos involucrados en una serie de comentarios y que gracias a estos se pueden determinar la utilidad de un material audiovisual publicado en la red. Este trabajo se enfoca en la clasificación de comentarios haciendo uso del algoritmo BERT clasificando videotutoriales basándonos en la clasificación de sus comentarios. Con los resultados obtenidos, además de las métricas, encontramos una motivación para un desarrollo futuro de la posible futura de la propuesta brindada en este Trabajo utilizando nuevos métodos de clasificación o incluso mejorando el propuesto.

Trabajos Futuros

Dentro de la elaboración de este trabajo existen variedad de ideas que consideramos pueden ser integradas en propuestas futuras que mejoren la precisión de modelo planteado en este escrito. Este trabajo se enfoca principalmente en la clasificación de videos sobre la base de sus comentarios, que colocan a los sentimientos en tres posibles clases. Las ideas que consideramos pueden ser agregadas en trabajos futuros son las siguientes:

- El uso de un dataset enfocado en comentarios neutrales, pues estos pueden parecer positivos al comienzo, pero su neutralidad puede afectar la clasificación de estos, dando resultados que puedan afectar la precisión del modelo.
- Enfoque del modelo en videos específicos, haciendo uso de términos especiales dentro de un contexto, por ejemplo: Iteraciones o referencias en un contexto de programación.
- El uso de más clases que puedan considerar los posibles giros que posean los comentarios neutrales.
- La forma en la que el modelo es construido puede ser diferente, pues en esta propuesta se hace uso de BERT, puede que se use otra herramienta para poder llevar a cabo la clasificación de comentarios de manera más efectiva y mejore las métricas, de la misma manera generando una matriz de confusión más grande pero más entendible y mucho más clara.

Referencias

[1] Melissa Carvalho Costa Alexandre Ashade Lassance Cunha and Marco Aurelio C. Pacheco. Sentiment analysis of youtube video comments using deep neural networks. In Lecture Notes in Computer Science, pages 561–570, 2019.

[2] Salman Aslam. Youtube by the numbers: Stats, demographics & fun facts. Omnicore, March 14, 2022, https://www.omnicoreagency.com/youtubestatistics/.

[3] Hanif Bhuiyan, Jinat Ara, Rajon Bardhan, and Md Rashedul Islam. Retrieving youtube video by sentiment analysis on user comment. In 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pages 474–478. IEEE, 2017.

[4] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. . O'Reilly Media, Inc.", 2009.

[5] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. KnowledgeBased Systems, 226:107134, 2021.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[7] Navoneel Chakrabarty. A machine learning approach to comment toxicity classification. In Computational intelligence in pattern recognition, pages 183–193. Springer, 2020.

[8] Jin Ding, Hailong Sun, Xu Wang, and Xudong Liu. Entity-level sentiment analysis of issue comments. In Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering, pages 7–13, 2018.

[9] ENOIT DURAND. 500+ programming ytb comments. Kaggle, https://www.kaggle.com/datasets/bdok /774-programming-ytb-commentsdataset/code?resource=download.

[10] Abbi Nizar Muhammad, Saiful Bukhori, and Priza Pandunata. Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. In 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pages 199–205. IEEE, 2019.

[11] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. Identifying toxicity within youtube video comment. In International conference on social computing, Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation, pages 214–223. Springer, 2019.

[12] Ayushka Tiwari Ritika Singh. Youtube comments sentiment analysis. International Journal of Scientific Research in Engineering and Management (IJSREM), 2021.

[13] Jalaj Thanaki. Python natural language processing. Packt Publishing Ltd, 2017.

Aplicación de Norma ISO 9241-11 para la Evaluación de la Usabilidad en Simuladores de Vuelo

Application of ISO 9241-11 Standard for the Evaluation of Usability in Flight Simulators

María Soledad Martínez
Dirección de Análisis Operativo. Fuerza Aérea, Córdoba, Argentina.
mariasolemartinez81@gmail.com
https://orcid.org/0000-0003-2346-9859

Daniel Ignacio Martínez
Universidad Tecnológica Nacional, Córdoba, Argentina.
danielignaciomartinez@gmail.com
https://orcid.org/0000-0001-6017-8132

Valeria Raquel Filoniuk
Dirección de Análisis Operativo. Fuerza Aérea, Córdoba, Argentina.
vfiloniuk@gmail.com
https://orcid.org/0000-0003-0614-3814

Gabriel Germán Chiappori
Dirección de Análisis Operativo. Fuerza Aérea, Córdoba, Argentina.
ggchiappori@gmail.com
https://orcid.org/0000-0003-4592-9342

Ana Claudia Diz
Dirección de Análisis Operativo. Fuerza Aérea, Córdoba, Argentina.
anaclaudiadiz@gmail.com
https://orcid.org/0000-0002-0585-860X

Silvia Edith Arias
Universidad Tecnológica Nacional, Córdoba, Argentina.
edith.edit@gmail.com
https://orcid.org/0000-0001-9695-2812

ARK: ark:/42411/s9/a68
PURL: 42411/s9/a68

RECIBIDO 17/08/2022 • ACEPTADO 15/09/2022 • PUBLICADO 30/09/2022

RESUMEN

Este artículo presenta la aplicación de la Norma ISO 9241-11 al software correspondiente a Simuladores de Vuelo de la Fuerza Aérea Argentina, con el fin de evaluar la usabilidad de dichos entrenadores. Cada organización y producto software son en general diferentes, es decir, no existe una prueba de usabilidad "única" que sea portable para aplicar a todos los "proyectos software". Empresas de gran prestigio, como Apple, Yahoo, Microsoft, entre otras, utilizan diferentes técnicas de usabilidad, en función de sus necesidades específicas. El principal objetivo de este trabajo es tomar como referencia la Norma ISO 9241-11 y adaptarla a las necesidades de la organización en cuestión, formulando e implementando métricas que ayuden a evaluar la usabilidad de manera objetiva y finalmente contrastar, analizar y reportar los resultados obtenidos en esta investigación aplicada.

Palabras claves: Métricas, Norma ISO 9241-11, Simuladores de Vuelo, Usabilidad.

ABSTRACT

This article presents the application of the ISO 9241-11 Standard to the Flight Simulators Software of the Air Force Argentine to evaluate the usability of these simulators. Each organization and software product are generally different, there is not a "unique" usability test that can be applied to all "software projects." Highly prestigious companies, such as Apple, Yahoo, and Microsoft, use different usability techniques depending on their specific needs. The main objective of this work is to take the ISO 9241-11 Standard as a reference and then adapt it to the organization´s needs, formulating and implementing metrics that help to evaluate usability objectively and finally contrast, analyze and report the results obtained in this applied research.

Keywords: *Metrics, Norma ISO 9241-11, Flight Simulators, Usability.*

INTRODUCCIÓN

El diseño de software centrado en el humano es un enfoque para el desarrollo de sistemas interactivos que tiene como objetivo hacer que los sistemas sean utilizables y útiles, centrándose en los usuarios, sus necesidades, requisitos, y aplicando factores humanos, ergonomía, conocimientos y técnicas de usabilidad. El objetivo de esta perspectiva es mejorar la efectividad y la eficiencia, el bienestar humano, la satisfacción del usuario, la accesibilidad y la sostenibilidad; y además contrarresta los posibles efectos adversos del uso en la salud humana, la seguridad y el rendimiento. [1]

El proceso de verificación y validación (V&V) aborda todas las fases del ciclo de vida del software, siendo utilizado para establecer si determinada etapa, tarea o producto, cumple con las necesidades del usuario y los requisitos establecidos para su desarrollo [2], V&V coadyuva al proceso de construcción proporcionando una valoración objetiva de los productos y los procesos que forman parte del ciclo de vida del desarrollo de software, es decir, garantiza que el producto final se ajuste a su respectiva especificación y que este cumple las expectativas del usuario. [3]

Las pruebas de software son parte de un proceso más amplio de verificación y validación de software, y se soporta en los estándares IEEE1008 e ISO / IEC 29119. Estas pruebas nacen por la necesidad de garantizar un producto de calidad, descubriendo defectos que podrían contener los programas antes de la implantación, y demostrar que un programa hace lo que se pretende que haga. Al verificar se realiza retroalimentación, es decir se vuelve a repasar el funcionamiento del plan en el pasado o en el presente para poder tomar acciones a la salida del feedback. Al validar se quiere tener la certeza que el sistema alcanzará el resultado definido en un plan a futuro y con posibilidad de que existan revalidaciones en el proceso.

La norma ISO/IEC 9241 está orientada hacia la calidad en usabilidad y ergonomía para productos y servicios en tecnología, tanto en software como en hardware, creada por la ISO (Organización Internacional de Normalización) y por la IEC (Comisión Electrotécnica Internacional). Estas normativas se sustentan en estándares como ISO/EEC 15288:2008 e ISO/EEC 12207:2008, que permiten aportar al software el concepto de calidad, estableciendo si los requisitos son correctos, completos, precisos, consistentes y verificables.

La norma ISO 9241 se enfoca en el diseño centrado al humano, y uno de los puntos claves es la usabilidad. La usabilidad se refiere a la capacidad de un software de ser comprendido, aprendido, usado y atractivo al usuario, en condiciones específicas de uso [4], [5].

La usabilidad es uno de los aspectos más importantes en los últimos años, y es una consecuencia del constante avance tecnológico y el deseo de ofrecer un producto que ayude a cumplir las metas del usuario. La norma 9241 establece el concepto de usabilidad [6] aplicado a sistemas interactivos [7], pero no es un proceso específico en la evaluación del diseño.

Las pruebas de usabilidad cumplen un rol fundamental en todo el proceso de V &V, siendo estas un atributo primordial de la calidad [8]. Sin embargo, los métodos de medición apropiados para evaluar la usabilidad no son obvios y son una preocupación constante para el personal involucrado en el desarrollo de un proyecto de software [9].

Cada empresa y producto son diferentes, motivo por el cuál no hay una prueba de usabilidad "única" que sea portable para aplicar a todas las organizaciones. Empresas de gran prestigio como Apple, MailChimp, Yahoo, DirecTV, Microsoft, Buffer, entre otras, utilizaron diferentes técnicas de usabilidad en función de sus necesidades específicas [10].

La evaluación de la usabilidad en simuladores de vuelos es una tarea compleja, donde no es suficiente que el software correspondiente a dicho dispositivo de entrenamiento funcione correctamente, sino también será necesario que el uso de éste sea satisfactorio a los pilotos. Si dicho dispositivo es percibido por los operadores como malo, deficiente o insatisfactorio, constituirá para ellos un mal sistema de adiestramiento, dificultándose en gran medida su capacitación mediante el mismo, razón por la cual su desarrollo sería en vano. Por tal motivo, la aceptación por parte del usuario será determinante para el éxito o fracaso de los simuladores de vuelo [11].

No obstante, la satisfacción del usuario es un indicador blando, con un marcado componente subjetivo, convirtiéndose su estimación en un desafió [12], ya que está más enfocada hacia las percepciones y actitudes de los usuarios, que hacia criterios concretos y objetivos [11].

En este trabajo se presenta la formulación e implementación de dos métricas de usabilidad, que *constituyen una medida de evaluación objetiva y de carácter supletoria a la satisfacción del usuario*, tomando como referencia los atributos de calidad que propone la Norma ISO 9241-11.

Materiales y métodos o Metodología computacional

Existen numerosos métodos para evaluar la usabilidad de un software. Cada una de ellos cuenta con sus propias ventajas y desventajas. Cada empresa u organización deberá seleccionar el método que más se adecué en función de sus necesidades específicas, teniendo en cuenta las características del sistema en cuestión o la etapa de desarrollo, entre otros [13]. En ocasiones resulta conveniente combinar estos métodos, con el fin de optimizar los resultados [10].

Según el Estándar ISO 9241-11, la usabilidad es entendida como "El grado en que un producto puede ser usado por usuarios específicos para lograr un objetivo con eficacia, eficiencia y satisfacción en un contexto de uso específico" [4], [8], [14], [15], [16], [17].

La eficacia puede ser obtenida a partir del porcentaje de tareas ejecutadas debidamente por el/los usuarios escogidos. En lo que respecta a la eficiencia, puede ser medida por medio del tiempo empleado para realizar las tareas establecidas por el evaluador.

En cuanto a la satisfacción, es definida como la capacidad del software para cumplir con las expectativas del usuario en un contexto de uso determinado. En consecuencia, la obtención de la satisfacción es un proceso con cierto grado de subjetividad, lo que no permite parametrizar cuantitativamente este atributo, algo que sí es posible con la eficacia y eficiencia [8].

Existen gran cantidad de métodos para el proceso de diseño de producto (PDP), la selección de cuál escoger depende muchas veces de las necesidades específicas del equipo de desarrollo y su contexto.

La investigación exploratoria se realizó mediante la búsqueda de artículos y bibliografía relacionada sobre las distintas técnicas existentes destinadas a evaluar la usabilidad, seleccionando para el desarrollo del objeto de estudio, la técnica más adecuada, con el objetivo de llegar a un resultado favorable en la aplicación de la misma al software de los simuladores de vuelo.

La metodología elegida para realizar pruebas de usabilidad en los referenciados ut-supra, se basa fundamentalmente en combinar una técnica cuantitativa con una técnica cualitativa, tomando como referencia la Norma ISO 9241-11, adaptándola a las necesidades específicas de la organización en cuestión. Ante ello, se *formularon e implementaron métricas correspondientes*

destinadas a la evaluación de la eficacia y eficiencia, acompañadas por la utilización de cuestionarios y observación directa, como medio para la valuación de la satisfacción del usuario.

Se ofrece en este marco a los usuarios del simulador, una metodología ágil de evaluación, basada en un concepto *disrruptivo* para la institución, de prestación y gestión de servicios, centrado en la mejora progresiva de procesos y en la interacción continua con los usuarios.

Con esta nueva visión se pretende ofrecer aumentos significativos en la precisión, alcance y cobertura en los simuladores, por parte de los pilotos.

Resultados y discusión

En esta sección se presentan los resultados del proceso de investigación, obtenidos a partir de la formulación e implementación de las métricas que a continuación se detallan en la figura 1. Los resultados alcanzados, previos y posteriores a la incorporación de dichas métricas, se presentan mediante dos momentos.

Al final de cada momento se presentan los aportes de la investigación general y la integración de los resultados, para luego dar paso a las conclusiones en el siguiente apartado.

Medición de la Efectividad: se evalúa que el usuario cumpla de forma correcta sus objetivos. Para esta evaluación se elaboró el siguiente indicador:

$$PTC = \frac{CTU*100}{CTDO} = \frac{\text{Cantidad Tareas que realizo el usuario para cumplir el objetivo de una prueba} *100}{\text{Cantidad Tareas Totales que debe realizar el usuario para cumplir el objetivo de una prueba}}$$

Si PTC>= 70%: Efectividad satisfactoria
Si PTC< 70%: Eficiencia no satisfactoria

Medición de la Eficiencia: se evalúa que el usuario cumpla con los tiempos promedios estimados para la realización de los ejercicios.

$$TR = \frac{TpoUT*100}{TEP} = \frac{\text{Tiempo requerido por el usuario para completar las tareas en la ejecución de la prueba} *100}{\text{Tiempo estándar de ejecución de esta prueba}}$$

Si TR>= 70%: Eficiencia satisfactoria
Si TR< 70%: Eficiencia no satisfactoria

Figura 1. Métrica para evaluar Efectividad y Eficiencia

En la Tabla 1, se detallan los resultados alcanzados, previos a la implementación de las métricas descriptas en la figura 1.

En este primer momento, el *único atributo considerado para la evaluación de la usabilidad se determina mediante la satisfacción del usuario.*

Tabla 1. Evaluación de la usabilidad mediante la satisfacción del usuario

Resultados obtenidos en un primer momento				
	Planeación	Ejecución	Análisis y Reporte	Resultado obtenido
Evaluación Satisfacción del Usuario	Se diseña un cuestionario destinado a evaluar la satisfacción del usuario.	Los usuarios completan los cuestionarios. Se analizan gestos, expresiones faciales y actitudes del usuario mientras interactúa con el simulador.	Se evalúa la satisfacción del usuario mediante cuestionarios y observación directa.	Subjetividad en las pruebas.

En la Tabla 2, se presentan los resultados obtenidos, posteriores a la implementación de las métricas mencionadas en la Figura 1.

En este segundo momento, la usabilidad es evaluada mediante la eficiencia, eficacia y satisfacción, tomando como referencia la Norma ISO 9241-11.

En ambos momentos, la evaluación se realiza tomando como base el proceso correspondiente de pruebas de usabilidad, que incluye las etapas de planeación, ejecución, análisis y reporte [15].

Tabla 2. Evaluación de la usabilidad mediante eficacia, eficiencia y satisfacción.

Resultados obtenidos en un segundo momento				
	Planeación	Ejecución	Análisis y Reporte	Resultado obtenido
Evaluación eficacia y eficiencia	Se diseña un ejercicio práctico de simulación a ejecutarse por los usuarios seleccionados. El rol de observador es designado a un miembro del equipo de Testing, cuya función es observar el comportamiento de los usuarios al momento de resolver las tareas y hacer anotaciones.	Los usuarios realizan el ejercicio de simulación propuesto por el Equipo de Testing. Este equipo, elabora un check list con los ítems que deben cumplirse, como así también los tiempos requeridos para la ejecución del mismo.	Se evalúa la eficacia y la eficiencia mediante las métricas formuladas, descriptas en la figura 1.	Confiabilidad, completitud y objetividad en las pruebas.
Evaluación Satisfacción del Usuario	Se diseña un cuestionario destinado a evaluar la satisfacción del usuario.	Los usuarios completan los cuestionarios. Se analizan gestos, expresiones faciales y actitudes del usuario mientras interactúa con el simulador.	Se evalúa la satisfacción del usuario mediante cuestionarios y observación directa.	

Los resultados obtenidos demuestran la importancia de combinar una técnica cuantitativa, como lo son la eficacia y la eficiencia, con otra cualitativa, como la satisfacción del usuario, en el proceso de pruebas de usabilidad, con el objetivo de obtener mejores resultados en las pruebas.

En un primer momento, la satisfacción del usuario fue tomada como única medida de prueba para la evaluación de la usabilidad, por considerarse a los usuarios finales los actores más importantes en el proceso de pruebas, quienes determinan la aceptación o rechazo del simulador, como instrumento de capacitación.

Sin embargo, los resultados correspondientes a los cuestionarios, preguntas abiertas, como así también gestos, expresiones faciales y actitudes frente al uso del simulador, en ocasiones difirió, entre un usuario y otro, en función de sus percepciones, experiencias y expectativas.

Cabe mentar, que un simulador de vuelo es un sistema que intenta replicar o simular la experiencia de pilotear una aeronave en particular, de la forma más realista posible. Para esto, se guardan los parámetros típicos del avión en vuelo, desde la velocidad de despegue, rutas, hasta posibles averías y accidentes.

Esto le permite al piloto formarse y al instructor, evaluarlo. En este contexto la aceptación del usuario no se refiere a la apreciación estética, sino más bien a la capacidad que tiene el dispositivo de comportarse de manera análoga a un avión real, tanto en software como en hardware.

Los resultados obtenidos corroboran que la denominada "satisfacción del usuario" es una medida subjetiva, ya que frente al uso del mismo dispositivo y bajo las mismas condiciones de uso, el software cumplió con las *expectativas* de los usuarios en distintas medidas, dependiendo en gran parte de sus *experiencias* previas frente al uso de otros simuladores. Las *percepciones* de similitud entre el simulador y la aeronave real, con respecto al hardware, varió en algunos casos, de un usuario a otro, valorando aspectos tales como, la sensibilidad de las palancas del tren de aterrizaje, de comando, como así también las pedaleras, entre otros.

Por lo expuesto anteriormente, en un segundo momento, se considera pertinente *complementar la satisfacción del usuario, con alguna medida de valoración objetiva, como lo son eficacia y eficiencia*, para la valoración de la usabilidad en los simuladores de vuelo, para lo cual se formularon e implementaron métricas en función de las necesidades específicas de la institución. *La implementación de las métricas, permitió cuantificar los resultados y obtener porcentajes correspondientes a la evaluación de la eficacia y la eficiencia. Estas métricas, junto a la satisfacción del usuario, resultaron ser el complemento ideal para la evaluación de la usabilidad de estos simuladores, logrando resultados que proporcionan un mayor nivel de confiabilidad, completitud y objetividad en las pruebas.*

Conclusiones

El propósito de esta investigación fue tomar como referencia la Norma ISO 9241-11 y adaptarla a las necesidades de nuestra Institución, formulando e implementando métricas que ayuden a optimizar las pruebas de usabilidad.

Los resultados obtenidos en esta investigación muestran que, a pesar que la satisfacción del usuario es determinante para el uso de los simuladores, la evaluación de la usabilidad mediante este atributo, como única medida de valoración, constituye una medida subjetiva, que depende en gran parte de las percepciones, actitudes y expectativas del usuario.

Este trabajo expone la formulación e implementación de dos métricas, destinadas a evaluar la eficacia y la eficiencia, siendo éstas medidas de evaluación objetivas y de carácter supletorias a la satisfacción del usuario, logrando de esta manera un mayor grado de confiabilidad, completitud y objetividad en las pruebas.

Agradecimientos

Quiero agradecer a todas las personas que nos apoyaron e hicieron posible que este trabajo se realice con éxito. A nuestra tutora, Esp. Ing. Silvia Arias, por su paciencia, tiempo dedicado y conocimientos brindados; a la profesora de inglés Gisela Codrington, por su contribución desinteresada en todo momento y a todo mi equipo de trabajo, que forman parte de mi labor diaria.

Referencias

[1] M. Mascheroni, C. L. Greiner, R. H. Petris, G. N. Dapozo and M. G. Estayno, "Calidad de software e ingeniería de usabilidad", in XIV Workshop de Investigadores en Ciencias de la Computación. La Plata, 2012, pp. 1-4. Available: http://sedici.unlp.edu.ar/handle/10915/19202

[2] Verificación y validación de software: una descripción general. Publicado en: IEEE Software, vol.6, no.1, pp.9-10, 1989. DOI: 10.1109/52.28119. Editor: IEEE

[3] M. D. Mosquera Pérez and L. M. Giraldo Castagno, "Formulación del modelo de gestión de procesos, bajo el enfoque de aseguramiento de la calidad, basado en el ciclo de mejora continua Phva de Edwards Deming, para el laboratorio de la industria académica en desarrollo

de software, para la facultad de ingeniería de la UCO," thesis, Universidad Católica de Oriente, Antioquia, 2019.

[4] N. Vázquez Callejón, "Análisis y desarrollo de heurísticas y guías de usabilidad de RESTFUL APIs y aplicación a un caso práctico", thesis, Universidade da Coruña, Coruña, 2020.

[5] D. A. Godoy, H. Bareiro, E. O. Sosa, E. Stoffel and G. Barros, "Usabilidad en simuladores web de redes de sensores inalámbricos," presented at XXI Workshop de Investigadores en Ciencias de la Computación, San Juan, 2019.

[6] W. Sánchez, "La usabilidad en ingeniería de software: definición y características," Rep. Investig, no.2, pp.7-21, Ago. 2011. Available: http://www.redicces.org.sv/jspui/handle/10972/1937

[7] SHARP, ROGERS and PREECE, "Interaction Design: Beyond Human-Computer Interaction," Wiley, 3rd

Edition. Available Biblioteca UGR: http://proquest.safaribooksonline.com/9780470665763

[8] D. M. Delgado Agudelo, D. F. Girón Timaná, G. E. Chanchí Golondrino and K. Márceles Villalba, "Estimación del atributo satisfacción en test de usuarios a partir del análisis de la expresión facial," I*ngenierías Universidad de Medellín*, vol.19, no.36, pp. 13-28, junio, 2019.

[9] J. R. Lewis, "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," International Journal of Human-Computer Interaction, vol.7, no.1, pp. 57-78, 1995.

[10] C. Bank, and J. Cao, *The Guide to Usability Testing*. (2014) [Online]. Available: https://www.inmagic.com

[11] C. R. Martín, "La satisfacción del usuario: Un concepto en alza," Facultad de Comunicación y Documentación y Servicio de Publicaciones de la Universidad de Murcia, vol. 3, pp. 139-153, 2000.

[12] A. De la Rosa Gómez, G. A. M. Díaz and S. X. M. Castillo, "Usabilidad y satisfacción de una aplicación móvil para el entrenamiento de competencias clínicas," Revista *Hamut'ay*, no.1, vol.7, pp. 48-59, abril 2020.

[13] G. G. Toribio, Y. P. Saldaña, J. J. H.Mora, M. J. S .Hernández, H. Bautista, C. A .Ordóñez and J. A. H. Alegría, "Medición de la usabilidad del diseño de interfaz de usuario con el

método de evaluación heurística: dos casos de estudio," *Revista Colombiana de Computación*, no.1, vol.20, pp.23-40, 2019.

[14] G. E. G. Chanchi, W. Y. M Campo and L. M. M. Sierra, Estudio del atributo satisfacción en pruebas de usabilidad, mediante técnicas de análisis de sentimientos," *Revista Ibérica de Sistemas e Tecnologías de Informação*, no.23, pp.340-352, 2019.

[15] S. V. Hernández and P. Chávez Lugo, Los Recursos Humanos como Factor Detonador de la Competitividad. Primer Edición. México: Editorial Ciempozuelos, 2019

[16] D. Albornoz, "Sistema software para la ejecución de pruebas de usabilidad bajo el enfoque de mouse tracking,". *TecnoLógicas.* vol.22, pp.19-31, 2019

[17] D. F. Ordoñez, and A. Bravo, "Aplicación de Heurísticas de Usabilidad de Nielsen sobre la Plataforma Moodle 2.8.3 + Build 20150225 de la Institución Universitaria Colegio Mayor Del Cauca," presented at II Congreso Internacional de Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil, 2018.

Uso de una herramienta de NLP aplicada a la detección del ciberacoso en Twitter

Use of an NLP tool applied to the detection of cyberbullying on Twitter

Jonathan Aguirre Soto
Universidad La Salle. Arequipa, Perú.
jaguirres@ulasalle.edu.pe

Hector Ávila Gonzales
Universidad La Salle. Arequipa, Perú.
havilag@ulasalle.edu.pe

Valeria Bravo Saines
Universidad La Salle. Arequipa, Perú.
vbravos@ulasalle.edu.pe

ARK: ark:/42411/s9/a65
PURL: 42411/s9/a65

RECIBIDO 10/07/2022 • ACEPTADO 28/08/2022 • PUBLICADO 30/09/2022

RESUMEN

En este documento se dará un breve resumen de como en la actualidad el constante desarrollo de la información y las tecnologías de comunicación (TICs) ha cambiado la interacción entre las personas hoy en día, por lo que las experiencias reales se han trasladado a un método virtualizado en este caso internet. Aunque las barreras de espacio-tiempo de la comunicación tradicional se han fragmentado, las relaciones sociales se han vuelto más fuertes, pero surgen nuevos problemas relacionados con diferentes conductas. El acoso, se define como un acto que amenaza el bienestar de una persona, y se convierte en ciberacoso cuando es realizado a través de internet generando a gran escala problemas de ansiedad, depresión e incluso el acto de suicidio y por lo cual es fundamental detectar a tiempo estos comportamientos malignos. Haremos uso de una herramienta de Procesamiento de Lenguaje Natural (NLP) utilizando Twitter como base para la extracción de las bases de conocimiento.

Palabras claves: Twitter, NLP, procesamiento de lenguaje natural, ciber-acoso, TICs.

ABSTRACT

This paper will briefly overview how the constant development of information and communication technologies (ICTs) has changed the interaction between people today. Real experiences have been transferred to a virtualized method, in this case, the internet. Although the space-time barriers of traditional communication have been fragmented, social relations have become more assertive, but new problems related to different behaviors arise. Bullying is defined as an act that threatens the well-being of a person and becomes cyberbullying when it is carried out over the internet, generating large-scale problems of anxiety, depression, and even the act of suicide, which is why it is essential to detect these malicious behaviors in time. We will use a Natural Language Processing (NLP) tool using Twitter as the basis for the extraction of knowledge bases.

Keywords: *Twitter, NLP, natural language processing, cyber-bullying, TICs.*

INTRODUCCIÓN

El ciberacoso es un acto reiterado que acosa, humilla o amenaza a las personas a través de sus ordenadores, teléfonos celulares, laptops, tabletas y otros dispositivos electrónicos, por otro lado también se incluyen sitios web que mantienen activas las redes sociales. Ciberacoso a través de internet utilizando las redes sociales, se ha vuelto más peligroso que el acoso tradicional porque tiene la capacidad de amplificar el daño y humillación a un grupo de personas que están conectadas en línea.

Muchas víctimas del acoso cibernético se sienten tristes, deprimidas, frustradas, incluso tienen pensamientos suicidas. Una encuesta realizada por UNICEF y el Ministerio de Comunicación e Información, en el Perú existe el cyberbullying y se encuentra en tasas de hasta un 40% entre los 13-15 años existe el bullying tradicional y de 11 a 14 años hay una tasa de incidentes de ciberacoso. Algunos de ellos podrían ser los acosadores; sin embargo, reconocen el peligro y los efectos negativos que llega a generar.

Por lo tanto, debería haber una forma de detectar a los principales actores del ciberacoso antes de que se den cuenta. reportar los daños que han causado en tiempo real. Por eso debe haber reglas para poder categorizar textos de ciberbullying. Por ello analizaremos la detección del ciberacoso mediante una herramienta de PNL usando un clasificador de texto.

Materiales y métodos o Metodología computacional

Motivación

El objetivo de este trabajo será desarrollar una herramienta para poder detectar las agresiones que se producen durante el uso de la aplicación utilizando lenguaje de procesamiento natural y un clasificador de texto a partir de datos de Twitter que nos dice que si es un comentario

sospechoso o es un comentario correcto mostrar a través de una clasificación binaria. Las preguntas que nuestro proyecto trata de responder son:

- ¿En qué dominio del conocimiento estamos trabajando?
- ¿Quiénes son los usuarios objetivo?
- ¿Por qué es interesante el tema propuesto?
- ¿Cuáles son las preguntas que nuestro proyecto de PNL trata de responder?

Trabajos Relacionados

1. **K-Nearest Neighbor (KNN):** el algoritmo K-Nearest Neighbor es el algoritmo de clasificación basado en instancias. Este algoritmo calcula la distancia (por ejemplo, la distancia euclidiana) o la distancia de similitud (por ejemplo, la similitud del coseno) entre el entrenamiento de datos y la prueba de datos.

   Los tweets de datos que han sido clasificados se transforman en un formato de grafico dirigido, con los usuarios de Twitter como nodos y las menciones en los tweets como bordes, como se muestra en la Figura [1]. El borde tiene un peso que representa cuantos tweets envía un usuario X, a otros usuarios Y o Z en un periodo de tiempo t. Además, contiene la lista de tweets clasificados del usuario X al usuario Y o Z.

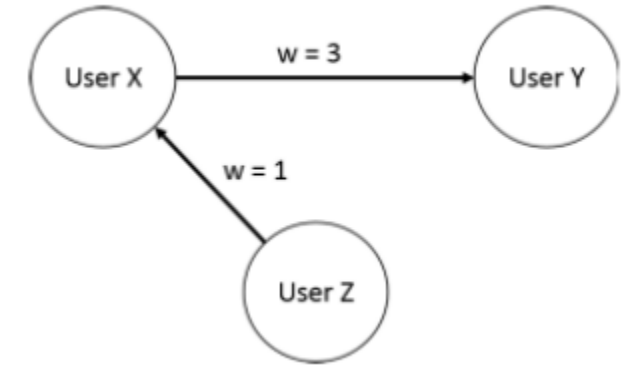

Figura 1. Visualización de datos gráficos de Twitter

2. **Ocho reglas generales para la extracción de características usando Sarna**

Estas reglas se pueden observar en la Figura [2] así como su definición a continuación:

a) El número de malas palabras en el tuit.
b) El número de palabras que muestran emociones negativas.

c) El número de palabras que muestran emoción positiva.
d) Combinación de un pronombre de primera persona, emoción negativa y combinación de un pronombre de segunda persona para capturar el ciberacoso.
e) Combinación de segundo pronombre con malas palabras para captar el ciberacoso.
f) Combinación de pronombres en primera persona, palabras que expresan emociones negativas y pronombres de tercera persona o nombres propios para capturar el ciberacoso.
g) Combinación de un pronombre de tercera persona o nombre propio con una mala palabra para capturar el acoso cibernético.
h) La combinación de enlace, blasfemia y pronombres también se utiliza para capturar el acoso cibernético.

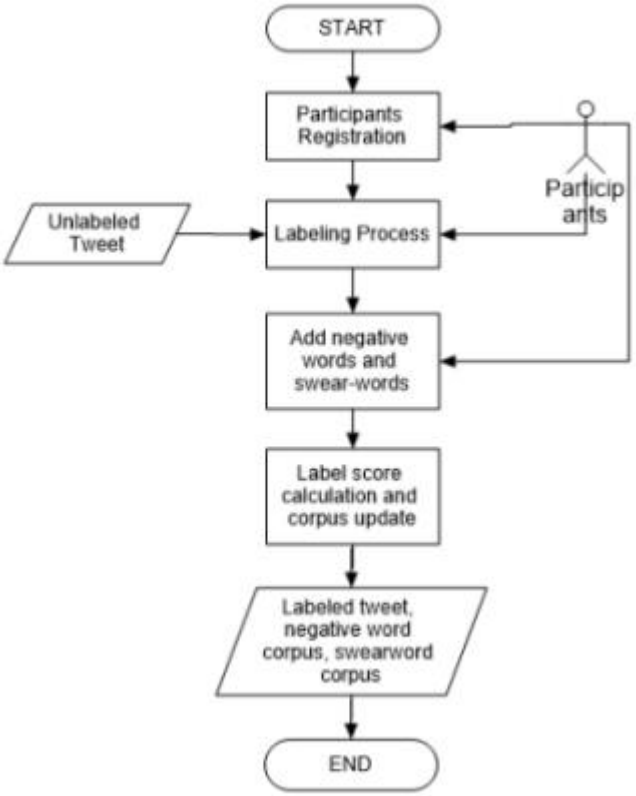

Figura 2. Sistema de etiquetado de datos

3. **spaCy:** Cada vez es más reconocido por procesar y examinar datos en PNL. Los datos textuales no estructurados se generan a gran escala y es fundamental para procesar y obtener información de datos no estructurados.

4. **Modelo BERT:** Actualmente es muy conocido puesto que Google está utilizando este modelo de lenguaje de procesamiento natural para mejorar sus búsquedas; En su lugar, lo hemos aplicado como un clasificador de texto para realizar el análisis de sentimiento.

5. **Tokenización BERT:** El comentario o Tweets seleccionados los empieza a tokenizar y nosotros lo hemos usado para obtener el conjunto de datos que contiene cada Tweet.

Propuesta

Para este trabajo, estamos utilizando un conjunto de datos de tweets relacionados con el ciberacoso, que habría que analizarlos en profundidad mediante la técnica de Sarna para poder clasificarlos de forma booleano; es decir, los que contengan 0 serán tuits de ciberacoso o de los que se pueda sospechar contienen cyberbullying, los que contengan 1 serán los tweets normales por lo que debemos realizar un sistema de etiquetado que podría utilizar diferentes personas que quisieran participar para etiquetar cada pieza de información, en cada uno de nuestros tuits en particular. Las personas pueden elegir una de las cuatro opciones. Opciones ofrecidas para un tweet. El propósito de usar cuatro opciones es permitir que las personas elijan una etiqueta para clasificar el tweet según su confiabilidad en un rango de 1 a 4 como se puede ver en la figura [3]. Esto se debe al nivel de confianza a la hora de elegir si el Tweet es ciberacoso o no.

Options	Weights	Descriptions
1	| −2 |	Very non-cyberbullying
2	| −1 |	Non-Bullying
3	1	Bullying
4	2	Very Cyberbullying

Figura 3. Opciones de etiquetado

1. **Etiquetado:** Las opciones 1 y 2 se utilizan para calcular la puntuación total sin ciberacoso. Para su parte, se utilizan las opciones 3 y 4 se utilizan para calcular la puntuación total de ciberacoso. Para diferenciar el grupo que acosa cibernéticamente y grupo que no acosa cibernéticamente, hay un signo menos en el peso del grupo que no acosa cibernéticamente.

   Finalmente, se obtuvieron los estadios de detección del ciberacoso. Nuestro principal objetivo es detectar cualquier situación de ciberacoso en Twitter; por lo que su funcionamiento parte de la adecuada formación de un algoritmo de aprendizaje supervisado. De ahí la importancia de proporcionar una base de conocimientos a partir de un análisis semántico y determinación de sentimientos para alcanzar la máxima tasa de precisión. Así, para llevar a cabo estas tres etapas ha sido necesario: (i) obtener la base de conocimientos, (ii) capacitación de modelos de aprendizaje supervisado, y (iii) implementación de estudios de casos.

2. **Obtención de la base de conocimientos:** Se debe establecer una base de conocimientos adecuada, comúnmente conocido como corpus, que interviene en la detección presuntiva del ciberacoso en lengua española como se puede ver en la figura [4]. Un corpus es un conjunto de palabras o frases que tienen previamente clasificados según diferentes intereses a través de etiquetas.

Pejorative word or insult	Synonym Ecuador
Animal, bestia	Huev*n
Mujer interesada	Grilla
Sinvergüenza	Caretuco
Imbécil	Careverg*
Bastardo	Hijo de put*
Temeroso	Ahuev*ado
Enojado	Arrech*
Fea/feo	Bagre
Feo/bajo status social	Batracio

Figura 4. Diez ejemplos de palabras insultantes y sus sinónimos en Ecuador.

3. **Entrenamiento de modelos de aprendizaje supervisado:** Se utilizarán tres métricas específicas de precisión.

   a) Puntuación de precisión, calculará la precisión del modelo de aprendizaje contando el número de muestras de ocurrencias que coinciden con el conjunto predeterminado de valores.
   b) Puntuación de precisión promedio, resumida en una curva de recuperación de precisión como la media de las precisiones alcanzadas en cada umbral.
   c) puntuación F1, se puede interpretar como un promedio ponderado de precisión y recuperación, donde una puntuación F1 alcanza su mejor valor en 1 y su peor puntuación en 0.

4. **Implementación de casos de estudio:** Existen tres tipos de análisis para la detección presuntiva de acoso cibernético.

   a) Análisis de oraciones y/o palabras, reflejará si el contenido de una oración o texto presumiblemente representa algún tipo de acoso.
   b) Análisis del perfil de usuario, presumiblemente determinando un porcentaje de acoso en base a un número predefinido de tweets históricos de un usuario específico. Para ello, introduzca el perfil en Formato de Twitter: @profildeuser.

Resultados y discusión

Análisis del modelo

El modelo utilizado fue BERT el cual nos ayudará a determinar si un Tweet es clasificado como cyberbullying (comentario agresivo, comentario insultante o comentario tóxico) o neutral (comentario que no es tóxico). La clasificación se desarrolla a partir de la propuesta realizada, ya que actualmente estamos clasificando los comentarios o Tweets en opciones de 1 (que es un comentario que contiene cyberbullying) y 0 (que es un comentario neutral), esto se puede ver en las Figuras [5] y [6] para que se vean las gráficas de clasificación.

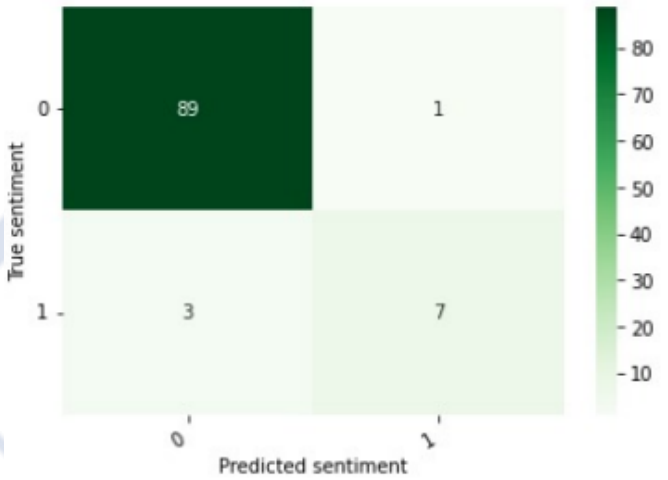

Figura 5. Clasificación de Tweets en opciones de 1 y 0.

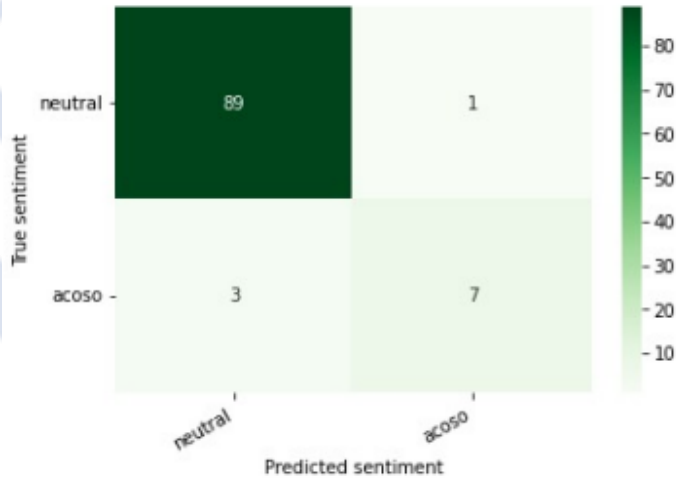

Figura 6. Clasificando los Tweets de acoso y los Tweets neutrales.

Análisis de clasificación

Sin embargo, para que BERT pudiera clasificar los Tweets, tuvo que pasar por un proceso de entrenamiento, pero ¿por qué entrenarlo? Es necesario hacerlo para que tenga un procesamiento más fluido, lo que hará de él un modelo más eficaz y eficiente por lo que los sentimientos que tiene que predecir serán más fácil de identificar, este entrenamiento se puede ver en la Figura [7], tuvieron una duración de 20 aproximadamente 40 minutos haciendo que la computadora esté activa durante al menos 2 horas para que pueda lograr el objetivo básico.

```
Epoch 1 de 5
------------------
/usr/local/lib/python3.7/dist-packages/transformers/tokenization_utils_base.py:2307: FutureWarning: The `pad_to_max_length`
  FutureWarning,
Entrenamiento: Loss: 0.3049762084893882, accuracy: 0.89875
Validación: Loss: 0.16842625822339738, accuracy: 0.95

Epoch 2 de 5
------------------
Entrenamiento: Loss: 0.2376163786649704, accuracy: 0.925
Validación: Loss: 0.29977081935586675, accuracy: 0.95

Epoch 3 de 5
------------------
Entrenamiento: Loss: 0.15388401919975878, accuracy: 0.95625
Validación: Loss: 0.2482876716447728, accuracy: 0.96

Epoch 4 de 5
------------------
Entrenamiento: Loss: 0.07530190275982022, accuracy: 0.975
Validación: Loss: 0.24376138745407974, accuracy: 0.95

Epoch 5 de 5
------------------
Entrenamiento: Loss: 0.04019912391435355, accuracy: 0.98875
Validación: Loss: 0.2275712515693158, accuracy: 0.96
```

Figura 7. Entrenamiento del Modelo Bert para clasificar los Tweets.

Resultados de entrenamiento

Después de entrenar al modelo, nos muestra los resultados de cómo realizó la clasificación tan pronto como con precisión n; es decir, qué tuits son realmente de acoso (etiqueta 1) y cuáles son tuits neutros (etiqueta 0), esto se puede observar en las Figuras [8] y [9] para dar una vista detallada de los resultados obtenidos durante el proceso y destacar que cada entrenamiento contiene a una parte de nuestro conjunto de datos inicial.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	90
1	0.88	0.70	0.78	10
accuracy			0.96	100
macro avg	0.92	0.84	0.88	100
weighted avg	0.96	0.96	0.96	100

Figura 8. Resultados del entrenamiento del Modelo BERT con etiquetas de 1 y 0.

	precision	recall	f1-score	support
neutral	0.97	0.99	0.98	90
acoso	0.88	0.70	0.78	10
accuracy			0.96	100
macro avg	0.92	0.84	0.88	100
weighted avg	0.96	0.96	0.96	100

Figura 9. Resultados del entrenamiento del Modelo BERT con etiquetas de acoso y neutral.

Conclusiones

Para finalizar, es importante mencionar que el ciberacoso es un problema actual que debemos solucionar inmediatamente, ya que muchas personas terminan perjudicadas por estos comentarios negativos en sus vidas, así que para llegar a su solución aplicaremos el procesamiento de lenguaje natural, el cual es una disciplina importante en nuestra actualidad porque gracias a su comprensión es posible identificar y analizar diferentes textos. Gracias a ello se puede aplicar en diferentes áreas para mayor beneficio, en este caso, el área que nos ocupa es ciberacoso en la aplicación de Twitter.

Como parte fundamental, este clasificador de texto nos ayudará a cumplir con el objetivo principal que es frenar el ciberacoso, implementando la aplicación del modelo BERT para lograr un resultado favorable y así poder frenar este problema denominado “ciberacoso”.

Referencias

[1] Gabriel A. Leon-Paredes, Wilson F. Palomeque-Leon, 2019. *Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language*. https://ieeexplore.ieee.org/abstract/document/8987684/

[2] Hani N., Dade N. *Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility*, 2018. https://ieeexplore.ieee.org/abstract/document/8350758/

[3] Monirah A. Al-Ajlan, Mourad Y. *Optimized Twitter Cyberbullying Detection based on Deep Learning*, 2018. https://ieeexplore.ieee.org/abstract/document/8593146

[4] Hoy interessa. *Cyberbullying in Peru stand at rates of up to forty per cent* https://gestion.pe/ tendencias/ciberbullying-peru-situa-tasas-40-140489-noticia/

[5] Orhan G. Yalcin. *Sentiment Analysis in 10 Minutes with BERT and TensorFlow* https:// towardsdatascience.com/sentiment-analysis-in-10-minutes-with-bert-and-hugging-face-294e8a04b6

[6] TensorFlow. *Classify text with BERT* https://www.tensorflow.org/text/tutorials/classify_text_with_bert

[7] Kaggle. Classified Tweets https://www.kaggle.com/datasets/munkialbright/classified-tweets

[8] GitHub. Cyberbullying Detection in Tweets https://github.com/apeksha104/ Cyberbullying-Detection-in-Tweets

[9] V Krithika, V Priya. A Detailed Survey On Cyberbullying in Social Networks https://ieeexplore. ieee.org/abstract/document/9077794

[10] Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu S. Choi, Byung-Won On Aggression detection through deep neural model on Twitter https://www.sciencedirect.com/science/ article/abs/pii/S0167739X19330717

[11] Bandeh A. Talpur, Declan Oaˆ€™Sullivan Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter https://www.mdpi.com/2227-9709/7/4/52

Revisión del proceso de mejora de software

Review of software process improvements

Diego Grell Casaverde Carpio
Universidad La Salle. Arequipa, Perú.
dcasaverdec@ulasalle.edu.pe

Jhonny Frans Gallegos Mendoza
Universidad La Salle. Arequipa, Perú.
jgallegosm@ulasalle.edu.pe

Jhoel Huallpar Dorado
Universidad La Salle. Arequipa, Perú.
jhuallpard@ulasalle.edu.pe

ARK: ark:/42411/s9/a70
PURL: 42411/s9/a70

RECIBIDO 03/06/2022 • ACEPTADO 24/07/2021 • PUBLICADO 30/09/2022

RESUMEN

A lo largo del tiempo han surgido diferentes modelos de mejoras de procesos para evaluar la calidad del software y aplicar mejoras en base a la evaluación, dentro los que más destacan están CMMI e ISO/IEC 15504 y MPS son los tres modelos principales para evaluación y mejora de procesos de software. Las diferencias que se abordan en el artículo son ventajas y niveles de madurez. CMMI se basa en las ideas de una madurez representada en un marco con de 5 niveles. ISO/IEC 15504, anteriormente conocida como SPICE. posee un marco de 6 niveles de madurez. MPS-BR está basado en CMMI con la diferencia de que posee 7 niveles de madurez. La principal diferencia entre CMMI y MPS-BR y la ISO 15504 es su orientación. Mientras CMMI está dirigido a grandes empresas, MPS-BR se enfoca en medianas y pequeñas empresas y la ISO 15004 está orientada a cualquier tipo de empresas, ya sea grandes empresas o las PyMES.

Palabras claves: Mejora de procesos, Procesos de software, CMMI, ISO ,MPS.Br.

ABSTRACT

Over time, different process improvement models have emerged to assess software quality and apply improvements based on the evaluation. Among the most outstanding are CMMI and ISO/IEC 15504, and MPS are the three main models for evaluating and improving software processes. The differences that are addressed in the article are benefits and maturity levels. CMMI is based on the ideas of evil reproduced in a framework with five levels. ISO/IEC 15504, formerly known as SPICE. It has a framework of 6 groups of maturity. MPS-BR is based on CMMI with the difference that it has seven maturity levels. The main difference between CMMI, MPS-BR, and ISO 15504 is their orientation. While CMMI is aimed at large companies, MPS-BR focuses on medium and small companies, and ISO 15004 is aimed at any company, whether large companies or PyMEs.

Keywords: *Process improvement, Software processes, CMMI, ISO, MPS.Br.*

INTRODUCCIÓN

Hoy en día las empresas que se encargan de desarrollar software buscan dos cosas, ganar más dinero en menos tiempo y maximizar la calidad del producto de software, esta premisa abarca aspectos muy importantes dentro de la calidad de software como disminuir costos, maximizar eficiencia del sistema, entre otros.

La mayoría de los proyectos de desarrollo de software enfrentan los siguientes problemas: Retraso en proyectos, sobrepasar el presupuesto y/o los clientes no están satisfechos con la calidad del software entregado. Esto es tan común que incluso tiene su propia denominación: crisis de software[1].

Hace algunos años se entendió que no había suficiente presupuesto para la resolución de problemas relacionados con el software [2] y entonces se centró más en la organización y cuestiones metodológicas.

Los procesos de software se aceptan como el área de ingeniería de software con más importancia durante la última década. Las investigaciones sobre la madurez del proceso de software proporcionaron información sobre las actividades del software e introdujo varios modelos de procesos de software que ayudaron a evaluar y mejorar tanto la capacidad del proceso de software como la madurez de organización productora de software.

La mejora del proceso de software busca mejorar o ampliar la forma en la que se lleva a cabo un proceso de la elaboración de un software, manteniendo la eficacia y la eficiencia del producto, El objetivo principal es analizar y definir cómo mejorar las prácticas de desarrollo de software dentro

de una empresa u organización. [3] Existen varios modelos de proceso, pero la evolución de estos procesos dejó tres frameworks conocidos como MOPROSOF CMMI y SPICE con sus revisiones más conocidas: MPS.Br CMMI e ISO/IEC 15504. Estos tres son los modelos más relevantes y los más importantes a nivel mundial.

MPS.Br es un programa para la mejora de los procesos de software desarrollo en el Brasil; Este programa se centra en mejorar la competitividad de las micro, pequeñas y medianas empresas de desarrollo de software, mejorando la calidad de los productos de software y sus servicios asociados, como en los procesos de producción y distribución de software.

Siendo su objetivo la mejora del proceso de software en algunos países en vías de desarrollo de Latinoamérica, con foco en las Micro, Pequeñas y Medianas empresas a un costo accesible.

Se escogió al estándar ISO/IEC 15504 porque se desempeña mediante la experimentación en la industria, además promueve la transferencia de tecnología de la evaluación de procesos de software.

CMMI-DEV se centra en prácticas para el correcto desarrollo de productos o servicios con una calidad estandarizada con el objetivo de lograr satisfacer las necesidades de los consumidores

La mayoría de proyectos de software existentes siempre tiene dificultades al escoger el proceso de mejora de software que se adecue a las necesidades del proyecto lo cual desemboca en que los proyectos no lleguen a culminar.

Por lo cual escogió este campo de investigación para lograr brindar la información necesaria en el proceso de selección. El propósito de este artículo es investigar cómo estos tres modelos están relacionados y se diferencian entre sí para lograr definir para qué proyecto están mejor enfocados y son más eficientes.

Materiales y métodos o Metodología computacional

El método que se utilizó para la elaboración del artículo, fue una exhaustiva búsqueda de un tema de interés, posteriormente se realizó la recolección de los diversos artículos, las cuales fueron extraídas de fuentes confiables y verídicas como: Google académico, Scielo y Redalyc.

Los criterios para la selección de un artículo, fueron las siguientes:

- Búsqueda de documentos acerca de la gestión y administración del agua potable.
- Realizar la búsqueda de documentos referentes con las mejoras de proceso de software.
- Artículos que aporten al trabajo de revisión.

- Revisar el contenido principal de los documentos preseleccionados.

Estos criterios utilizados para la selección de la información, ayudará a obtener una revisión más confiable y con problemas actuales, y de esta manera nuestro artículo de revisión cumplirá con el propósito planteado anteriormente.

Resultados y discusión

ISO/IEC 15504 es una norma que propone un modelo para evaluar la capacidad en los procesos de desarrollo de un determinado producto de software.

Esta norma está basada en los siguientes objetivos:

- Proponer y desarrollar un estándar que se encargue de evaluar los procesos de software.
- Evaluar el desempeño del desarrollo de software mediante la experimentación de la industria.
- Promover la transferencia tecnológica de en análisis y evaluación de procesos de software en esta industria a nivel mundial.

En la actualidad la industria del software ha tenido grandes avances por lo que es necesario que se impongan nuevos estándares de calidad para la certificación de procesos de desarrollo, de modo que se acrediten a estas organizaciones para que brinden un mejor servicio a un mercado que cada día es más grande, mucho más internacional y competitivo.

Capacitación organizacional (OT) tiene como objetivo desarrollar el conocimiento y las habilidades de los empleados. Su objetivo es permitir a los empleados llevar a cabo sus funciones de manera eficiente y eficaz. Tiene como objetivo habilitar empleados para cumplir con los objetivos comerciales de la organización y satisfacer los requisitos de entrenamiento táctico [4]. Las Medidas acompañados con el primer CMMI específico es un marco para evaluar y mejorar los proyectos de software que se desarrollan por el Instituto de Ingeniería de Software (SEI) en Carnegie Universidad de Mellon en los Estados Unidos.

Preguntas de objetivos

Se aplicó el paradigma de métricas (GQM) en la organización área de procesos de formación en CMMI. Se aplicó para definir las medidas de objetivos específicos y sus prácticas específicas.

MPS.BR es un programa para la mejora de los procesos de software de las pequeñas y medianas empresas de desarrollo en Brasil. Los modelos planteados por MPS.Br están basados en conceptos

de madurez y capacidad de proceso para la evaluación y mejora de la calidad y productividad de productos de software y servicios asociados. Este programa se centra en mejorar la competitividad de las micro, pequeñas y medianas empresas de desarrollo de software, mejorando la calidad de los productos de software y sus servicios asociados, así como en los procesos de producción y distribución de software.

Tabla 1. Ventajas de los tres principales modelos de mejora de proceso ISO 15504, CMMI, MPS.Br

VENTAJAS		
ISO 15504	CMMI-DEV	MPS.Br
Se adapta a la forma de trabajo de cada organización pues no reemplaza la forma de trabajar.	Al tener un objetivo en común esto hace que la comunicación interna y externa mejore.	• Utilización más eficiente de los recursos.
No establece procesos obligatorios para que una empresa los ejecute.	Aumenta la calidad de los productos y servicios con lo cual reduce los tiempos de entrega.	Es flexible y se adapta a los nuevos proyectos.
No esta centrado en el cumplimiento del proceso, más bien se enfoca en la realización y gestión de los procesos	Disminución de gasto en el presupuesto.	Los modelos de madurez establecen una hoja de ruta evolutiva y gradual para la implementación de mejoras en los procesos.
Tiene la capacidad de ser certificable por todas las entidades que generan certificación del panorama social.	Es un modelo que cuenta con muchos años de experiencia ya que cuenta .	Es frecuentemente utilizado como criterio de selección y calificación de proveedores por parte de grandes empresas públicas y privadas.

Tabla 2. Niveles de madurez de los tres principales modelos de mejora de proceso ISO 15504, CMMI, MPS.Br

NIVELES DE MADUREZ		
ISO 15504	CMMI-DEV	MPS.Br
Esta norma presenta 6 niveles que se detallan a continuación [5]: • Nivel 0 - Incompleto: El proceso no se encuentra implementado. • Nivel 1 - Realizado: Se evalúa el proceso a nivel corporativo, se encuentra evidencia de la realización del proceso. • Nivel 2 - Gestionado: Los procesos se encuentran mapeados y se gestionan de modo que se establezcan, controlen y mantengan. • Nivel 3 - Establecido: Se somete el proceso adaptado a un proceso estándar. • Nivel 4 - Predecible: El proceso es evaluado y gestionado usando diversas técnicas cuantitativas. • Nivel 5 - Optimizado: Se busca que el proceso mejore continuamente de modo que se cumplan los objetivos del negocio actuales y los propuestos.	Esta norma posee 5 niveles de madurez. [6] • Nivel 1 - Inicial: Proceso es informal y Ad Hoc • Nivel 2 - Gestionado: Básica para la Gestión de Proyectos Administra los requisitos, tiene planificación del proyecto, medición y análisis • Nivel 3 - Definido: Estandarización de procesos. Tiene desarrollo de requisitos, solución técnica, integra los productos, verificación, validación, enfoque del proceso. organizacional y gestión integrada de proyecto. • Nivel 4 - Cuantitativamente gestionado: Rendimiento del proceso organizacional gestión y gestión de proyectos cuantitativos. • Nivel 5 - Optimizado: Mejora continua de los procesos, innovación en organización y despliegue. análisis causal y resolución	Este modelo consta de 7 niveles de madurez: [7] • Nivel A - En optimización: Fase inicial del proyecto, proceso informal. • Nivel B - Gestionado Cuantitativamente: Se desarrolla la gestión de proyectos. • Nivel C - Definido: Se define la gestión de decisiones y la gestión de riesgos. • Nivel D - Ampliamente Definido: Se da el desarrollo de requisitos y la integración del producto. • Nivel E - Parcialmente Definido: Se da la evaluación y mejora del proceso organizacional. • Nivel F - Gestionado: Se da la gestión de portafolios y el aseguramiento de la calidad. • Nivel G - Parcialmente Gestionado: Se da el cambio del proceso y la mejora continua.

Conclusiones

Este artículo contribuye a diferenciar los diferentes procesos de software con:

- Establecimiento de niveles de madurez ISO/IEC 15504, CMMI y MPS
- Establecimiento de principales ventajas de ISO/IEC 15504, CMMI y MPS
- Definición de cada uno de los procesos ISO/IEC 15504, CMMI y MPS

CMMI puede ser representado como una opción a la norma ISO 15504, interpretado como un sistema de evaluación de madurez de procesos. CMMI en la actualidad es mucho más conocido internacionalmente que ISO 15504 y MPS ya que tiene mayor presencia en EEUU por lo que resulta obligatorio para las empresas.

Los 3 modelos presentan una estructura parecida bajo niveles de madurez, en el caso de MPS tiene 7 niveles y en el caso de CMMI y ISO 15504 tienen 6 niveles, además en los tres casos tienen áreas de proceso muy similares que definen el estándar de calidad para la organización.

Los modelos estudiados en la mejora de procesos, se adecuan a las necesidades de cada proyecto, cada proceso de mejora se puede aplicar según las necesidades de la organización, CMMI y la ISO 15504 son los modelos de mejora de procesos más conocido y aplicado actualmente, pero el modelo MPS.Br va en ascenso principalmente en Latinoamérica. Agradecimientos (Opcional) "A mis docentes y en especial a nuestro tutor Yasiel Pérez por su ayuda, paciencia y dedicación Agradecerle también a toda nuestra familia por darnos ánimo durante este proceso. A nuestros amigos de toda la vida que nos acompañan desde siempre.

Referencias

[1] F. Rabbanikhah, A. M. Jaghagh, R. M. Gholizadeh, S. Sabouri, and S. Alirezaei, "Analyzing effective factors in efficiency of organizational trainings (A Case Study: Employees of Ministry of Health and Medical Education)," International Journal of Humanities and Cultural Studies (IJHCS) ISSN 2356-5926, pp. 2136-2154, 2016

[2] C. P. Team, Capability Maturity Model® Integration for Development Version 1.3 (Software Engineering Institute). 2010.

[3] F. J. Pino, F. García, M. Piattini, Software process improvement in small and medium software enterprises: A systematic review. Software Quality Journal. 16, 237–261 (2008).

[4] A. L. Mesquida, A. Mas, Implementing information security best practices on software lifecycle processes: The ISO/IEC 15504 Security Extension. Computers and Security. 48, 19–34 (2015).

[5] www.calidadygestion.com, ISO 15504. Calidad y Gestión (2021), p. 9.

[6]Tutorialspoint.com. 2022. SEI CMMI - Niveles de Madurez. [online] Available at: [Accessed 8 July 2022].

[7] Alvarado, R., Delgado, L., & Dávila, A. (2012, July). Mapeo y evaluación de la cobertura de los procesos de MPS. Br a los procesos de la categoría de Operación de MoProSoft. In Anais do XI Simpósio Brasileiro de Qualidade de Software (pp. 158-172). SBC.

Predicción del nivel de obesidad en personas usando el modelo de árbol de decisión

Prediction of the level of obesity in people using the decision tree model

Renato Eduardo Delgado Huacallo
Universidad Nacional de San Agustín.
Arequipa, Perú.
rdelgadoh@unsa.edu.pe
https://orcid.org/0000-0002-6222-5294

Christian Ilachoque Hanccoccallo
Universidad Nacional de San Agustín.
Arequipa, Perú.
cilachoque@unsa.edu.pe
https://orcid.org/0000-0003-4468-6713

Felman Luque Sanabria
Universidad Nacional de San Agustín.
Arequipa, Perú.
fluques@unsa.edu.pe
https://orcid.org/0000-0001-6322-0228

Jose Maykol Paniura Huamani
Universidad Nacional de San Agustín.
Arequipa, Perú.
jpaniura@unsa.edu.pe
https://orcid.org/0000-0002-3031-9317

ARK: ark:/42411/s9/a71
PURL: 42411/s9/a71

RECIBIDO 20/06/2022 • ACEPTADO 31/07/2022 • PUBLICADO 30/09/2022

RESUMEN

La obesidad es un problema para la salud pública que afecta a la población mundial es por eso que el presente trabajo está orientado a presentar una solución informática para la estimación y predicción de niveles de obesidad haciendo posible que una persona pueda conocer su estado físico actual, para esto se usó un dataset de personas con obesidad de distintos países como Perú, México y Colombia basándose en sus hábitos alimenticios y su condición física, creando con todos estos datos un árbol de decisión.

Palabras claves: Obesidad, árbol de decisión, nivel de obesidad.

ABSTRACT

Obesity is a public health problem that affects the world population, that is why the present work is oriented to present a computer solution for the estimation and prediction of obesity levels, making it possible for a person to know their current physical condition for this we used a dataset

of people with obesity from different countries like Peru, Mexico and Colombia based on their eating habits and their physical condition, creating a decision tree with all these data.

Keywords: *Obesity, decision tree, level of obesity.*

INTRODUCCIÓN

Es un hecho en el mundo que el problema de obesidad va en aumento, muchos años atrás está tan solo era un número pequeño dentro de la población mundial, pero este problema en la actualidad va en crecimiento, lo cual puede presentar un gran problema para la salud pública de infantes y adultos, cuyas consecuencias propias de un actuar negligente, representa pérdidas sustanciales, en razón a que la obesidad implica un estado vulnerable hacia otras enfermedades como puede ser el cáncer o problemas cardiovasculares que afectan al corazón [1,2].

Según la Organización Mundial de la Salud (OMS) [3] desde 1975 los problemas de obesidad presentaron un incremento cercano al triple de su media habitual de aquel entonces, trasladándonos al año 2016 que servirá como punto de referencia y del cual destacan los siguientes datos: El 39% de adultos, considera aquellos de 18 años a más tenían sobrepeso y el 13% eran obesos; más de 340 millones de niños y adolescentes, considera aquellos de 5 a 18 años contaban con sobrepeso u obesidad y alrededor de 41 millones de niños menores a los 5 años de edad contaban con sobrepeso u obesidad. Manifestando entonces, la poca o nula preocupación frente al problema, que en determinados casos deviene en consecuencias irreversibles y más aún teniendo en cuenta que el tratamiento para restablecer la salud y estado físico no es de periodicidad inmediata a lo cual este lapso de tiempo será de vital importancia para los individuos pertinentes al caso.

Con el fin de adecuar cuidadosamente la información sobre este tema se ha tomando en cuenta algunos trabajos relacionados, de los que se pueden destacar los siguientes: El primero, hace un estudio en la población española de adultos, donde se evalúa la prevalencia de obesidad considerando mediciones antropométricas individuales, factores sociodemográficos, consumo alimentario, actividad física, estilos de vida y problemas de salud. Con el que se llegó a la conclusión de que existía una prevalencia de obesidad alta en varones con mayor edad, y este factor presenta relación inversa con el nivel socioeconómico. Además de que la probabilidad mínima de obesidad estaba relacionada a las personas con estilos de vida que incluían la actividad física y un sedentarismo moderado.[4]

Como segundo trabajo, también realizado en España, en esta ocasión el estudio fue realizado a una población infantil. En este caso se evalúa la estimación de la prevalencia de sobrepeso en niños de entre 2 y 14 años, tomando en cuenta el Índice de Masa Corporal (IMC) y propiamente

dicho la edad y sexo de la población de niños y niñas. De este trabajo se obtuvo que la prevalencia de sobrepeso y obesidad alta era mayor en los varones [5].

Un tercer trabajo, nos describe un proyecto de software realizado en Arequipa en el cual se analiza, diseña e implementa un modelo de minería de datos con datos recolectados de diferentes colegios del Perú para estimar en nivel de obesidad de una persona mediante el IMC, en el cual se hace uso del algoritmo de árboles de decisión para poder predecir un resultado de acuerdo a datos ingresados por el usuario en este caso el estudiante. Se usan varios algoritmos para esta tarea, tales como el J48, Multilayer Perceptrion, ForestPA, NaiveBayes, BayestNet asi obteniendo como resultado el mejor algoritmo que es J48 con una precisión del 94.38% [6].

Como último trabajo relacionado tenemos una investigación en la cual se identifican las variables más influyentes en determinar el grado de obesidad por medio de técnicas de minería de datos, en esta se tiene 16 variables independientes y una variable dependiente la cual se identifica como grado de obesidad, se hace uso del algoritmo J48 y otras técnicas de inteligencia artificial para poder cumplir el objetivo, los resultados de la investigación muestran que las variables independientes más influyentes son el género, estatura, peso y el IMC, así también se obtiene que por el algoritmo J48 se obtiene un éxito superior al 97% mediante validación cruzada [7].

En este trabajo se pretende utilizar un dataset recuperado del repositorio de Machine Learning UCI [8], el cual presenta datos para la estimación de distintos niveles de obesidad en personas de los países de México, Perú y Colombia, basados en sus hábitos alimenticios y condición física. Por todo lo descrito anteriormente, el presente trabajo tiene como objetivo principal presentar una solución informática para la estimación y/o predicción de niveles de obesidad de las personas que deseen conocer su estado físico; haciendo uso de la data set mencionado anteriormente y utilizando como técnica de análisis predictivo un árbol de decisión, el cual está implementado en el lenguaje de programación de Python.

Trabajos relacionados

En el trabajo de Moral et.al. [9], desarrollan una estrategia para evaluar variables que puedan influir en la aparición o evolución de la Diabetes tipo 2. La información almacenada de los pacientes incluye historias clínicas y datos de laboratorio; dicha información se encuentra almacenada en distintas bases de datos. La técnica utilizada es del Random Forest, el cual crea árboles de decisión para la clasificación; concluyendo que su mejor modelo obtenido es un Random Forest con 40 árboles y una profundidad de 5 para cada uno, listando además las variables más importantes para la predicción.

Otro trabajo que se puede rescatar es el de Fierro et.al. [10], en donde realizan un análisis de tres modelos predictivos que están basados en Machine Learning, para obtener la predicción de la tendencia de los jóvenes al alcoholismo. Dentro del dataset se tiene información del estado familiar, lugar de vivienda de un joven, entre otros como variables de entrada. Como resultado de su análisis obtuvieron que el modelo con mayor precisión fue el modelo de Regresión Lineal, quedando por debajo el modelo KNN y el Árbol de decisión.

Materiales y métodos o Metodología computacional

Como se mencionó anteriormente, se recuperaron los datos acerca de la estimación de niveles de obesidad del repositorio de Machine Learning UCI [8], presentando información recolectada de personas de los países de Colombia, México y Perú; este dataset se compone de distintas variables de entrada (como el género, edad, altura, peso, historia familiar con sobrepeso, entre otras) y una variable de salida que representa el nivel de obesidad. A continuación, se hará una breve descripción de las variables de entrada restantes:

- FAVC: Consumo frecuente de alimentos ricos en calorías
- FCVC: Frecuencia de consumo de verduras
- NCP: Número de comidas principales
- CAEC: Consumo de alimentos entre comidas
- SMOKE: Fuma (Si o No)
- CH2O: Consumo de agua diario
- SCC: Seguimiento del consumo de calorías
- FAF: Frecuencia de actividad física
- TUE: Tiempo usando dispositivos tecnológicos
- CALC: Consumo de alcohol
- MTRANS: Transporte usado

Entre las herramientas que se utilizaron, se tiene en primer lugar Google Collab, el cual nos permite ejecutar código de Python en el navegador para distintos usos, tales como IA, análisis de datos entre otros. Este servicio no requiere instalación y nos brinda la posibilidad de acceso a recursos computacionales sin costo.

Nuestra otra herramienta usada es Scikit-learn es cual es una biblioteca, considerada como la más útil y sólida para lo que se refiere aprendizaje automático en Python, se basa principalmente en Numpy, SciPy y Matplolib. También se ha hecho uso de Pandas, usado para tareas destinadas a ciencias de datos y aprendizaje automático, está construido sobre el paquete Numpy.

Como se indicó en la introducción, el modelo que se planea usar es el árbol de decisión; para ello se necesita tener el dataset en formato CSV y guardarlo en Google Drive para acceder desde el código Python. Antes de realizar el modelo, se necesita hacer una limpieza de datos para detectar datos anómalos, para ello se lee el archivo CSV y se grafican histogramas de los datos que pertenecen a las variables que se consideran necesarias a analizar, en este caso se verifican las variables de edad, altura y peso.

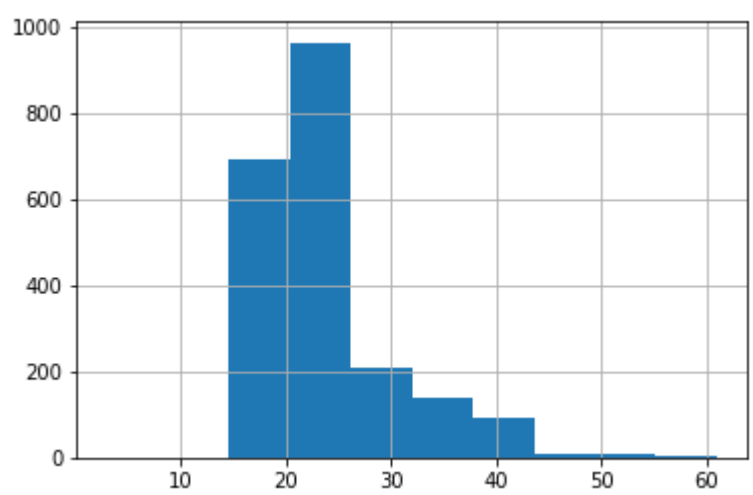

Figura 1. Histograma de Edad

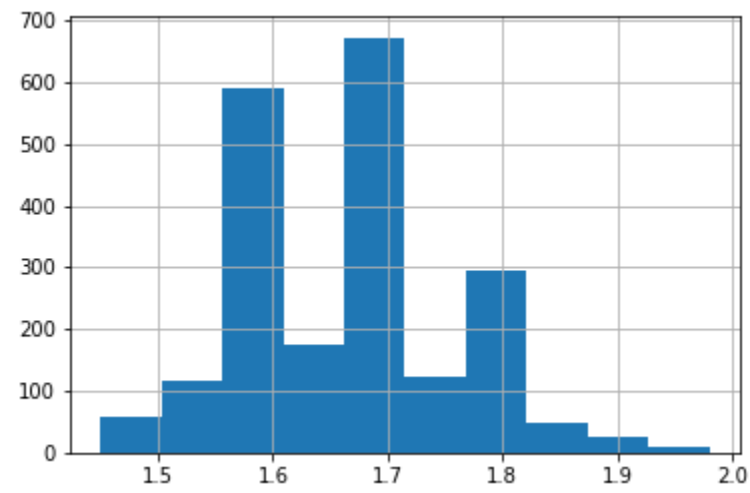

Figura 2. Histograma de Altura

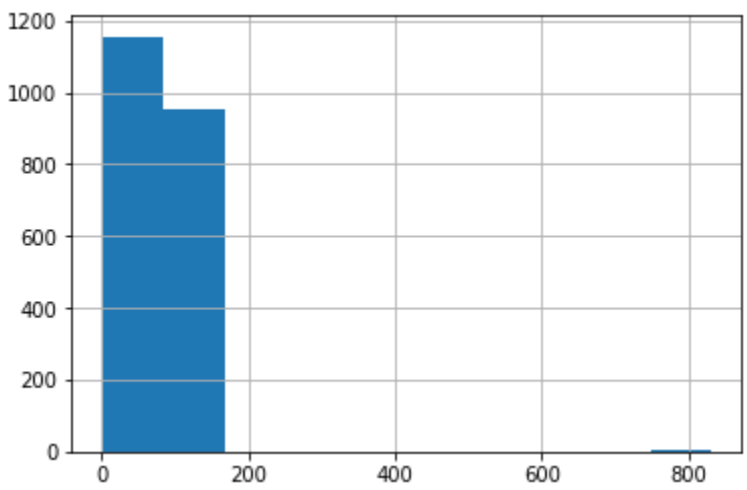

Figura 3. Histograma de Peso

En dichos histogramas se pueden observar algunos datos que están muy alejados al conjunto con mayor cantidad de datos, por ende, se procede a filtrar el dataset, usando el siguiente código:

```
filtered_dataset = dataset[(dataset['Age'] > 15) & (dataset['Age'] < 45) & (dataset['Height']<1.88) & (dataset['Weight']<155)]
```

Figura 4. Código para filtrar el dataset

Una vez realizada la filtración del dataset, se procede a graficar de nuevo los histogramas, en este caso con todas las variables de entrada del dataset.

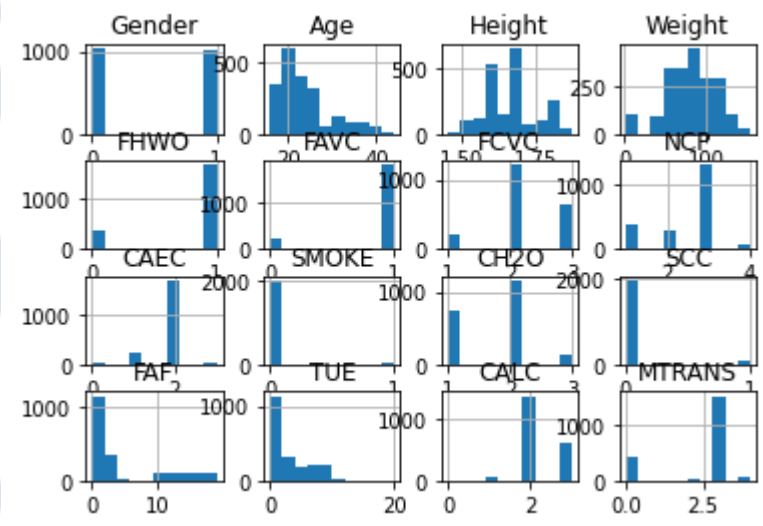

Figura 5. Histogramas después de realizar el filtro

Después se utiliza un 20% de los datos para la realización de pruebas, y el 80% restante de los datos para realizar el entrenamiento. Ahora se procede con la creación del modelo de árbol de decisión, en esta oportunidad se plantea desarrollar un árbol con profundidad de 4. En adición se realiza la validación del entrenamiento usando la matriz de confusión y la métrica de exactitud del modelo; finalmente se procede a crear el árbol de decisión.

Resultados y discusión

En la validación del entrenamiento, primero se realizó la matriz de confusión dando como resultado lo siguiente:

```
matriz = confusion_matrix(Y_test, Y_pred)
print("Matriz de confusion")
print(matriz)
```

```
Matriz de confusion
[[33 12  0  3  1  0  0]
 [ 3 45  0  0  0  1  2]
 [ 0  0 46  8  0  1 15]
 [ 0  0  1 60  0  0  0]
 [ 0  0  0  0 57  0  0]
 [ 0 15  0  2  1 26 22]
 [ 0  2 14  4  0  6 30]]
```

Figura 6. Matriz de confusión

En cuanto a la exactitud del modelo, se obtuvo el valor de 0,724390243902439, lo cual indica que el modelo no es tan exacto, pero se consideraría aceptable.

```
accuracy_score = accuracy_score(Y_test, Y_pred)
print('Exactitud del modelo:')
print(accuracy_score)
```

```
Exactitud del modelo:
0.724390243902439
```

Figura 7. Resultado exactitud del modelo

Para finalizar, se presenta el árbol de decisión que se generó en Python.

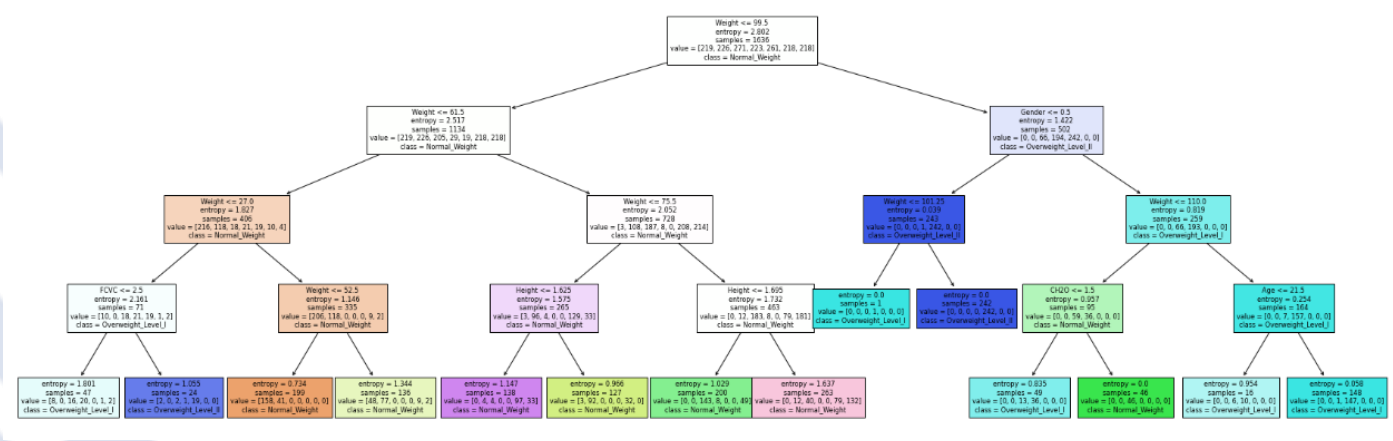

Figura 8. Árbol de decisión generado

Conclusiones

Los resultados obtenidos a lo largo del proyecto en el que se ha utilizado un árbol de decisiones para aplicación práctica a un Dataset sobre nivel de obesidad, resultan en una exactitud de 0.7243 que depende de la cantidad de datos de entrenamiento que se proporciona; aunque no es un valor bastante exacto, es un resultado aceptable.

Para trabajos futuros se tiene pensado implementar y estudiar casos similares para realizar una comparación entre diferentes modelos de Machine Learning para poder evidenciar diferentes perspectivas de solución, así como analizar ventajas y desventajas que se encuentren para cada modelo.

Referencias

[1] M. Malo Serrano, N. Castillo M. y D. Pajita D., "La obesidad en el mundo", Anales de la Facultad de Medicina, vol. 78, n.º 2, p. 67, julio de 2017. Accedido el 21 de junio de 2022. [En línea]. Disponible: https://doi.org/10.15381/anales.v78i2.13213

[2] L. M. T. Garcia, R. F. Hunter, K. Haye, C. D. Economos y A. C. King, "Un marco conceptual orientado a la acción para soluciones sistémicas de prevención de la obesidad infantil en Latinoamérica y en las poblaciones latinas de Estados Unidos", Obesity

Reviews, vol. 22, S5, octubre de 2021. Accedido el 21 de junio de 2022. [En línea]. Disponible: https://doi.org/10.1111/obr.13354

[3] Organización Mundial de la Salud (OMS), Nota descriptiva N°311 junio de 2016. Disponible en: http://www.who.int/mediacentre/factsheets/ fs311/es/

[4] Pérez-Rodrigo, C., Hervás Bárbara, G., Gianzo Citores, M. y Aranceta-Bartrina, J. (2021). Prevalencia de obesidad y factores de riesgo cardiovascular asociados en la población general española: estudio ENPE. Revista Española de Cardiología. https://doi.org/10.1016/j.recesp.2020.12.013

[5] Lasarte-Velillas, J. J., Hernández-Aguilar, M. T., Martínez-Boyero, T., Soria-Cabeza, G., Soria-Ruiz, D., Bastarós-García, J. C., Gil-Hernández, I., Pastor-Arilla, C. y Lasarte-Sanz, I. (2015). Estimación de la prevalencia de sobrepeso y obesidad infantil en un sector sanitario de Zaragoza utilizando diferentes estándares de crecimiento. Anales de Pediatría, 82(3), 152–158. https://doi.org/10.1016/j.anpedi.2014.03.005

[6] M. Ticona, "Sistema Para la Predicción de Obesidad en la Adolescencia Utilizando Técnicas de Minería de Datos", Universidad Católica de Santa María, Arequipa, 2018. Accedido el 21 de junio de 2022. [En línea]. Disponible en: http://tesis.ucsm.edu.pe/repositorio/handle/UCSM/8305

[7] O. D. Castrillón, "Las variables más influyentes en la obesidad: un análisis desde la minería de datos", Información tecnológica, vol. 32, n.º 6, pp. 123–132, diciembre de 2021. Accedido el 23 de junio de 2022. [En línea]. Disponible en: https://doi.org/10.4067/s0718-07642021000600123.

[8] F. Mendoza Palechor y A. de la Hoz Manotas. (2019). UCI Machine Learning Repository: Estimation of obesity levels based on eating habits and physical condition Data Set. [En línea] Available: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+. [Accedido: Jun 21, 2022]

[9] MORAL, D. R., & GARCIA, L. C. ANALISIS PREDICTIVO EN DIABETES TIPO 2 USANDO ESTRUCTURAS BIG DATA A. RODRIGUEZ 1, 2, V. SUAREZ-ULLOA1, C. TILVE ALVAREZ1, P. PUIG GALLEGO1, A. SOTO GONZALEZ1.

[10] Fierro, F. S., Castañeda, J., & Revelo-Aldás, M. (2022). Modelos predictivos para la estimación de adolescentes con tendencia al alcoholismo. AXIOMA, 1(26), 74-79.

Revisión de los avances y cambios en ciberseguridad en el Perú, para una transformación digital

Review of advances and changes in cybersecurity in Peru, for a digital transformation

Edwin Daniel Leon Gutierrez
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
edwin.leon@unjbg.edu.pe
https://orcid.org/0000-0002-2519-1785

Cynthia Mayumi Tesillo Gomez
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
cynthia.tesillo@unjbg.edu.pe
https://orcid.org/0000-0002-1769-9845

Yuri Alexander Escobar Arcaya
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
yuri.escobar@unjbg.edu.pe
https://orcid.org/0000-0001-5739-3050

Luis Antonio Godoy Montoya
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
luis.godoy@unjbg.edu.pe
https://orcid.org/0000-0001-8860-8843

ARK: ark:/42411/s9/a62
PURL: 42411/s9/a62

RECIBIDO 29/05/2022 • ACEPTADO 05/07/2022 • PUBLICADO 30/09/2022

RESUMEN

El presente trabajo tiene por objetivo explorar publicaciones donde ha sido tratado el tema de ciberseguridad en el Perú. Para ello, se han revisado 12 artículos originales en relación a la temática, publicados en los últimos 5 años. Este artículo cuenta con un claro objetivo descriptivo, exploratorio.

Palabras claves: Ciberseguridad, ciberataques, Estándar, ISO, NTP.

ABSTRACT

The objective of this work is to explore publications where the topic of cybersecurity in Peru has been treated. For this, 12 original articles have been reviewed in relation to the subject, published in the last 5 years. This article has a clear descriptive, exploratory objective.

Keywords: *Cybersecurity, Cyberattack, Standard, ISO, NTP.*

INTRODUCCIÓN

La irrupción de nuevas tecnologías, la proliferación de modernos dispositivos inteligentes, dio origen a una transformación digital no planificada en muchos casos, incentivada en los dos últimos años por el confinamiento al que la población se ha visto sometida por la pandemia originada por el COVID-19. Si bien la situación representa una oportunidad esto también trae consigo algunos riesgos asociados al uso de tecnologías que hasta hace poco no tenían mucha difusión y junto con las amenazas se hace necesario una adecuada planificación de la gestión de las mismas, sobre todo en un entorno cibernético.

El uso masivo de nuevas tecnologías en el campo de las TI, y la sofisticación continua, está originando que los riesgos se tornen más peligrosos y se diversifiquen. La gestión de la información ha sido considerada como uno de los más preciados activos en toda organización desde inicios de los sistemas de información por ser base para la toma de decisiones con impacto directo o indirecto en las personas [1].

En la actualidad los medios digitales están vinculados a potenciales vulnerabilidades, las cuales de no ser controladas/minimizadas facilitaran el accionar de agentes externos a los sistemas, con el peligro de alteración, robo, secuestro de información o recursos digitales [2]. Es ante estos riesgos latentes que las estrategias en tema de Ciberseguridad se vuelven en una necesidad que debe asumir una nación a fin de garantizar el bienestar de la población [2]. Para [3] en el Perú no existe por parte del estado una real toma de conciencia respecto a los posibles daños que podrían generar a las empresas los ciberataques. La importancia que han ido alcanzando los ciberataques queda identificada por [4] que para el año de la publicación asignaba al Perú el quinto lugar en cuanto a ciberataques registrados.

En cuanto a protección de datos personales, se tiende a dar prioridad a aquellos campos en donde el uso masivo de información, ha ido de la mano con el incremento de servicios ofertados online, sin embargo debe de prestarse también atención a aquellas áreas en donde el resguardo de la privacidad de la información resulta crítica aun cuando no se vincule a un uso masivo de la misma tal es el caso de información sensible generada por usuarios en la actividad privada tales como médicos, abogados, médicos cirujanos plásticos, etc [5]. Si bien el estado peruano ha dado pasos hacia un ordenamiento en temas de seguridad informática (ciberseguridad, seguridad de la información, etc) por medio de la implementación de normas, también es correcto decir que son pocos los avances en los aspectos de organización y capacitación [6]. Se llama en [7] a reflexionar respecto a dar la debida importancia a invertir tanto en tecnología como en recursos humanos, tal es así que se menciona que uno de los mayores riesgos para las entidades bancarias es el sabotaje efectuado por un insider, empleado de la misma organización. La necesidad de que el estado implemente estrategias nacionales de ciberseguridad no deben de limitarse a garantizar la seguridad de los ciudadanos y las infraestructuras, debe también incluir la instauración de un ecosistema que permita la cooperación público - privada [8]. Respecto a los rápidos cambios

tecnológicos a los que nos vemos enfrentados, una pregunta toma forma en [9] ¿Serán capaces de adaptarse rápidamente tanto las personas como las organizaciones a los cambios tecnológicos de los que somos testigos día tras día?, una respuesta negativa puede darnos una idea de lo vulnerables que serán ante las amenazas informáticas. Hemos pasado en poco tiempo de la euforia por los avances tecnológicos revolucionando la vida moderna a la preocupación por el desarrollo de riesgos potenciales a los que sin embargo no se les da la debida importancia [10].

Las PYMES no están ajenas a los abruptos cambios en temas de tecnología y son estas quienes puede ver en ellas una oportunidad de mejora, así como una fuente de potenciales riesgos sobre todo en época de pandemia en donde han sido duramente golpeadas financieramente, por lo que el tema de ciberseguridad se torna relevante para enfrentar vulnerabilidades que se pueden generar en el ciberespacio y afectar la continuidad del negocio [11]. Pero existen sectores altamente vulnerables que conforme van adoptando tecnologías IOT, van convirtiéndose en posibles blancos de ciberataques, tal es el caso de empresas dedicadas a servicios tales como tratamiento/potabilización de agua, en las cuales el alto grado de automatización supone un gran desafío para la implementación de sistemas de ciberseguridad [12]. En [13] el año 2019 se indicaba que únicamente el 9% de las empresas peruanas estaban aptas para detectar un ciberataque a tiempo, esta cifra nos da una idea de lo lejano que está que el estado en su conjunto pueda articular acciones conjuntas (sector público y privado) en temas de ciberseguridad y respuesta ante ciberataques. En cuanto a la "información", en [1] se indica que es cambiante y por tanto la protección de la misma adquiere esa característica. En el presente trabajo se presenta una revisión de publicaciones que abordan el tema de ciberseguridad en el Perú, para lo cual se ha realizado una búsqueda de material bibliográfico que permita tener un primer acercamiento a los avances que se realizaron en ciberseguridad en el Perú. Este artículo tiene por objetivo explorar los avances que se realizaron en ciberseguridad en el Perú a través de la revisión de papers publicados en los últimos 5 años.

Materiales y métodos

El presente trabajo es descriptivo exploratorio, para la recolección de información, se han utilizado unos criterios de búsqueda y de inclusión para seleccionar los artículos que finalmente han formado parte de la revisión. Ambos criterios se describen a continuación.

1. Búsqueda en bases de datos bibliográficas.
   Para realizar la búsqueda bibliográfica se realizó una búsqueda en diferentes bases de datos (Google académico, Dialnet, isoc).
   Las palabras clave utilizadas fueron: ciberseguridad, Perú, normas iso.

2. Criterios de inclusión.

Una vez obtenidos los resultados de las búsquedas a través de las técnicas anteriores, los artículos se pasaron por un filtro y solo se aceptó que formaran parte de la revisión aquellos que cumplían con los siguientes criterios:

- Tener acceso al texto completo del artículo científico.
- Excluir los artículos donde el tema de ciberseguridad fuese tratado secundariamente.
- Los documentos deben tener fecha de publicación mayor o igual al 2017.

Resultados y discusión

Después de una depuración de los 18 artículos, la base de datos quedó constituida por 12 artículos. A continuación, se muestra la distribución de artículos por año de publicación y tipo de artículo.

Tabla 1. Artículos por año de publicación

Año de publicación	N° de Artículos
2017	1
2018	4
2019	2
2020	2
2021	2
2022	1
Total	12

Fuente: Elaboración propia

Tabla 2. Artículos por tipo

Tipo de artículo	N°
Artículo Científico	8
Tesis de grado. Maestro	2
Tesis de grado. Licenciatura	1
Tesis de grado. Bachiller	1
Total	12

Fuente Elaboración propia

A continuación, se presentan los artículos seleccionados.

Título: "GESTIÓN DE LA CIBERSEGURIDAD Y PREVENCIÓN DE LOS ATAQUES CIBERNÉTICOS EN LAS PYMES DEL PERÚ, 2016"

Autor: ANTONIO INOGUCHI ROJAS, ERIKA LIZET MACHA MORENO - 2017

En el trabajo de investigación se aborda la situación de las PYMES del Perú ante los desafíos que supone la ciberseguridad, tema que hasta no hace mucho no representaba un foco de interés, sin embargo la masificación de los servicio online que han ido adoptando las PYMES las obliga a replantearse seriamente los riesgos que suponen estas nuevas tecnologías en el ámbito de protección de datos y la seguridad de los sistemas de información., teniendo en cuenta que al referirse a la data informática, se considera toda la información virtual almacenada y disponible en la red privada, siendo este recurso fundamental y vital para que las pymes funcionen correctamente y alcancen los objetivos propuestos. Los autores logran identificar (en la PYME objeto del trabajo) la falta de visión respecto a seguridad de la información, sobre todo en el tema de ciberseguridad, ya que la pérdida de información o manipulación por personas ajenas a la empresa conlleva a resultados adversos para la empresa misma, a tal punto de correr el riesgo de quiebra. Con los resultados obtenidos en la investigación, se proponen recomendaciones, indicando una propuesta para gestión y prevención de seguridad informática, la cual podrá ser

aplicable para la mayoría de pymes de diferentes rubros o giros de negocio, el único requisito es que la pyme se proponga implementar la propuesta de seguridad informática resultante.

Título: "CIBERSEGURIDAD EN LA INFRAESTRUCTURA CRÍTICA MEDIANTE EL SISTEMA SCADA EN PLANTA DE TRATAMIENTO DE AGUA DE LIMA"

Autor: André FERREIRA ALVES MACHADO, Lizet CACHO DE LA CRUZ. - 2018

En el trabajo se aborda el análisis de vulnerabilidades en un sector muy crítico como es el de prestación de servicios básicos específicamente abastecimiento de agua potable. Sectores claves como el que se trata, se respaldan en una amplia gama de sistemas y recursos informáticos para su funcionamiento continuo, confiable y efectivo. Estos sectores son conocidos como infraestructuras críticas. La protección de este tipo de infraestructura ha tomado más relevancia en los últimos años, por el gran impacto sobre la comunidad que supone los riesgos a los que está sometida, lo cual ha motivado a los Estados a generar acciones para garantizar su seguridad. Por otro lado, los cambios permanentes de las tecnologías hacen necesario no solo un profundo trabajo de articulación entre diversos actores sino la evaluación constante de distintos escenarios de compromiso, así como la adopción de medidas preventivas y correctivas para minimizar cualquier impacto de un ataque cibernético sobre los servicios esenciales. adicionalmente el trabajo define y caracteriza a las infraestructuras críticas, presentando la descripción de un conjunto particular de sistemas denominados SCADA o sistemas de control y adquisición de datos, así también describe casos de estudio de vulnerabilidades de dichos sistemas y alega sobre su implementación en la planta potabilizadora de agua en el Perú (Sedapal).

Título: "Análisis de la preparación de las organizaciones Mapfre Perú Seguros y Kallpa Corredora de Seguros ante las amenazas de seguridad de la información en el medio empresarial y que podrían impactar en sus operaciones de negocio".

Autor: Beteta Lazarte, Juan Enrique, Narva De la Cruz, Miluska de Jesús-2018

Trabajo en el que se analiza el sector de seguros privados, como el objetivo de evaluar cómo están preparadas Mapfre Perú Seguros y Kallpa corredora de seguros, dos empresas del mismo rubro pero con distinta capacidad financiera, ante las amenazas de seguridad de la información

que podrían impactar en sus operaciones de negocio, teniendo como finalidad el proponer una guía base de controles para mitigar los riesgos.

Título: "La ciberseguridad y el contexto actual"

Autor: Roberto Vizcardo Benavides - 2018

La guerra mundial en el ciberespacio está teniendo lugar. Ejércitos de hackers, espías informáticos y ciberdelincuentes conforman las fuerzas contrincantes; no hay distinción, los adversarios son naciones desarrolladas o en vías de desarrollo, así como grupos e individuos al margen de la ley. Generalmente no hay víctimas mortales ni heridos; pero los daños económicos son inconmensurables. Naciones Unidas y la OEA en particular, han tomado el reto de enfrentar esta nueva amenaza mediante la defensa cooperativa; sin embargo, en el ámbito regional, es poco lo que se ha hecho.

Título: "Propuesta de implementación de un modelo de gestión de ciberseguridad para el centro de operaciones de seguridad (SOC) de una empresa de telecomunicaciones".

Autor: Vilcarromero Zubiate, Ladi Lizeth; Vilchez Linares, Evit - 2018

En el trabajo, los autores proponen un modelo de gestión de ciberseguridad a una empresa de telecomunicaciones, sobre la base de una adecuada gestión del riesgo y la medición de controles según un nivel de madurez, dada la importancia del área de desenvolvimiento de la misma que es considerada muy crítica. Sectores como el que se abordan representan para la seguridad nacional y económica de los países, un factor importante del cual depende el funcionamiento confiable de su infraestructura crítica. Las amenazas de ciberseguridad explotan la creciente complejidad de dichos sistemas, colocando la economía, la seguridad pública y la salud en riesgo. Al igual que el riesgo financiero y de reputación, el riesgo de ciberseguridad afecta a los objetivos estratégicos de una empresa. Así mismo, la información se ha convertido en uno de los activos más importantes para cualquier organización, y el aseguramiento de la misma como un punto primordial para lograr ventajas competitivas y generación del valor, basando en el adecuado resguardo de la Confidencialidad, Disponibilidad e Integridad de la Información.

Título: "Problemática en ciberseguridad como protección de sistemas informáticos y redes sociales en el Perú y en el Mundo".

Autor: Alexis Enrique Poma Vargas - 2019

La seguridad en las redes sociales y sistemas de información son puntos de interés que se analizan en el trabajo, en el cual el autor se plantea investigar si la Ciberseguridad protege los sistemas informáticos y redes sociales en el Perú y el mundo, puesto que existen lugares tales como empresas y entidades propensas a todo tipo de ataques cibernéticos realizados por hackers, quienes sustraen información valiosa de estas.

Dicha investigación es cuantitativa, basada en análisis documentario, buscó información local, nacional e internacional de fuentes confiables, que pudiesen aportar alcances estadísticos y casos de protección a medios informáticos, a fin de proporcionar conocimientos sobre vulnerabilidad de sistemas por no contar con los recursos que permitan salvaguardar datos reservados, así como, las soluciones respectivas al caso.

Los resultados fueron favorables, en el sentido que se pudo apreciar que, mediante la adopción de medidas rígidas, los sistemas informáticos; así como las redes sociales, se vuelven potencialmente seguros en un 70%. En tal sentido se concluye que la ciberseguridad como protección de medios informáticos potencializa las buenas prácticas en las empresas y protege la información.

Título: "LATINOAMÉRICA ¿CÓMO ESTÁ AVANZANDO LA CIBERSEGURIDAD EN EL PERÚ? BREVE APROXIMACIÓN AL MARCO NORMATIVO".

Autor: Viviana García - 2019

Este artículo busca ejemplificar el marco normativo que se viene desarrollando en el país con relación a la ciberseguridad y mostrar que, si bien no es aparente, sí existe una preocupación por esta materia, dado que el Perú no puede estar ajeno al gran reto que representa para las organizaciones el proceso de transformación digital y que conlleva una necesaria reflexión respecto de las amenazas latentes por la potencial vulneración a la seguridad de los sistemas de información.

Título: "Estrategias integradas de ciberseguridad para el fortalecimiento de la seguridad nacional"

Autor: Juan Fernando Ormachea Montes - 2020

En el trabajo se establece como objetivo proponer estrategias integradas de ciberseguridad necesarias para fortalecer la seguridad nacional del Perú. Se evalúa las estrategias y políticas actuales utilizadas en el ámbito internacional para contrarrestar las ciber amenazas, así como las estrategias de ciberseguridad implementadas específicamente por los Países Bajos, EE. UU., España y Perú. Como resultado de la investigación se encontró que, en los indicadores referidos a cooperación regional, bilateral y multilateral, el Perú ha manifestado comportamientos disímiles; además, el Estado y la sociedad peruana aún transitan por los enfoques de la concientización y del desarrollo de las capacidades cibernéticas militares, como indicadores prevalentes en el diseño de las políticas nacionales de ciberseguridad. Por ello, se concluyó que la ciberseguridad constituye un compromiso social que demanda articulación entre el sector público y el sector privado, lo que en el Perú aún no se concreta; en consecuencia, el diseño de la Estrategia Nacional de Ciberseguridad del Perú constituye una necesidad que demanda ser satisfecha.

Título: "La seguridad de la información en la administración pública"

Autor: Kadú Josep Altamirano-de-la-Borda - 2020

En el trabajo se aborda la importancia de la información pública, la cual es producto de la administración y transformación de otra información que tiene un efecto directo en la ciudadanía, por lo cual debe ser protegida asegurando su confidencialidad, integridad y disponibilidad, teniendo siempre presentes los principios del derecho al acceso a la información de los ciudadanos. Si bien en el Perú se han dado los pasos adecuados para implementar los sistemas de gestión de seguridad de la información en el ámbito público, solo el 6 % de los organismos públicos ha cumplido con implementar lo dispuesto por la normativa vigente. Por lo cual se hace necesario que el Estado redoble los esfuerzos para proteger su información a través de la implementación de sus sistemas de gestión de seguridad de la información, ya que los nuevos escenarios referentes a la tecnología, la información y la interacción entre Estado y ciudadanía así lo requieren.

Título: "Análisis preliminar de la ciberseguridad asociada al sistema financiero en algunos países de Latinoamérica y la contribución de la informática forense".

Autor:Mayra A. Arévalo Álvarez, Daniel Andrey Hernández Ladino. -2021

En el trabajo los autores por medio de una investigación en algunos de los países Latinoamérica se sumergen en el sistema financiero bancario con el objetivo de identificar si los bancos se han visto expuestos a algún tipo de amenaza en años recientes asociados con su auge en la nube, que medidas de prevención han tomado o han buscado implementar para contrarrestarlas y cuál es la contribución que ofrece o puede ofrecer la informática forense para ayudar a esclarecer procesos de investigación principalmente relacionados con delitos financieros.

Título: Ciberseguridad y protección de datos personales en el Perú

Autor:José Álvaro Quiroga León - 2021

El trabajo realiza un análisis del marco legal en torno a la política de privacidad, así como el Reglamento de la Ley de Protección de Datos Personales.

Título: Marco de referencia "HOGO" para ciberseguridad en PyMES basado en ISO 27002 y 27032

Autor: Carlos Francisco Cruzado Puente de la Vega, Liset Sulay Rodriguez Baca - 2022

El trabajo resalta la importancia de desarrollar el marco referencial "HOGO" basado en las buenas prácticas del ISO 27002 y los controles de seguridad del ISO 27032 para la ciberseguridad en las PyMES. Los resultados de la investigación muestran los beneficios de la implementación del marco referencial "HOGO" en las PyMES, aplicando buenas prácticas relacionadas a la seguridad en internet, de las infraestructuras críticas para la información, seguridad de las redes y seguridad de la información. A medida que las tecnologías de información y comunicación se van empoderando en las organizaciones, se va generando una necesidad de protección del activo más importante, la información ante la necesidad de lograr protección de ataques en el ciberespacio.

Conclusiones

En el presente trabajo se realizó una revisión exploratoria de varias publicaciones que abordan el tema de ciberseguridad en el Perú, evidenciando que el interés en dicha área ha ido incrementando con el transcurso del tiempo. En los textos explorados los autores presentan varias perspectivas en cuanto a los avances en el tema de ciberseguridad, pasando por la comparación en el área de inversiones en el contexto de Latinoamérica, la situación del Perú respecto a los ataques cibernéticos, la implementación de normas que buscan establecer un ordenamiento tanto para el ámbito estatal como privado. Así también se identifica la falta de concientización y organización que son importantes para poder dar soporte a la implementación del marco normativo, y la importancia en invertir no solo en tecnología sino también en recursos humanos a fin de poder asegurar el minimizar todos los riesgos potenciales a los que se exponen las organizaciones públicas y privadas en temas de ciberseguridad producto del auge de las nuevas tecnologías y su masificación. La contribución del presente trabajo radica en que se presenta una revisión exploratoria de las publicaciones que en tema de ciberseguridad se han realizado en los últimos 5 años lo cual permite establecer una línea basal sobre la cual poder realizar otras investigaciones en los próximos años. Se sugiere como trabajo futuro el poder realizar una revisión con mayor profundidad en este tema dado que el continuo desarrollo de las TIC.

Referencias

[1] Kadú Josep Altamirano-de-la-Borda. La seguridad de la información en la administración pública Enlace: https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/13917/Altamirano_La-seguridad-de-la-informaci%C3%B3n-en-la-administraci%C3%B3n-p%C3%BAblica.pdf?sequence=1&isAllowed=y

[2] Freddy Linares. Vulnerabilidad en el sector público y la urgencia de pensar en ciberseguridad.(2022) Enlace: https://ciup.up.edu.pe/analisis/vulnerabilidad-en-sector-publico-la-urgencia-de-pensar-ciberseguridad/

[3] GARCÍA, V. (2019). ¿ CÓMO ESTÁ AVANZANDO LA CIBERSEGURIDAD EN EL PERÚ? BREVE APROXIMACIÓN AL MARCO NORMATIVO. Actualidad Jurídica (1578-956X), (52). Enlace: https://www.uria.com/es/publicaciones/6687-como-esta-avanzando-la-ciberseguridad-en-el-peru-breve-aproximacion-al-marco-n

[4] Inoguchi Rojas, A., & Macha Moreno, E. L. (2017). Gestión de la ciberseguridad y prevención de los ataques cibernéticos en las PYMES del Perú, 2016. Enlace: https://repositorio.usil.edu.pe/items/9449a061-bfd2-4ecc-8cf1-770fba7cee45/full

[5] León, J. Á. Q. (2021). Ciberseguridad y protección de datos personales en el Perú. Advocatus, (039), 15-21. Enlace: https://revistas.ulima.edu.pe/index.php/Advocatus/article/view/5114

[6] Vilcarromero Zubiate, L. L., & Vilchez Linares, E. (2018). Propuesta de implementación de un modelo de gestión de ciberseguridad para el centro de operaciones de seguridad (SOC) de una empresa de telecomunicaciones. Enlace: https://repositorioacademico.upc.edu.pe/handle/10757/624832

[7] Álvarez, M. A. A., & Ladino, D. A. H. (2021). Análisis preliminar de la ciberseguridad asociada al sistema financiero en algunos países de Latinoamérica y la contribución de la informática forense. Cuaderno de investigaciones: semilleros andina, 1(14). Enlace: https://revia.areandina.edu.co/index.php/vbn/article/download/1950/1873

[8] Ormachea Montes, J. F. Integrated cybersecurity strategies for strengthening national security. Enlace: https://www.recide.caen.edu.pe/index.php/recide/article/view/36

[9] Beteta Lazarte, J. E., & Narva De la Cruz, M. D. J. Análisis de la preparación de las organizaciones Mapfre Perú Seguros y Kallpa Corredora de Seguros ante las amenazas de seguridad de la información en el medio empresarial y que podrían impactar en sus operaciones de negocio.

[10] Benavides, R. V. (2018). La ciberseguridad y el contexto actual. Pensamiento Conjunto, 6(2), 9-9. Enlace: http://www.pensamientoconjunto.com.pe/index.php/PC/article/view/82

[11] Cruzado Puente de la Vega, C. F., & Rodriguez Baca, L. S. (2022). Marco de referencia "HOGO" para ciberseguridad en PyMES basado en ISO 27002 y 27032. Enlace: https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/5200/Carlos_Tesis_Maestro_2022.pdf?sequence=1&isAllowed=y

[12] Machado, F. A. (2018). Ciberseguridad en la infraestructura crítica mediante el sistema SCADA en planta de tratamiento de agua de Lima. Revista Escuela de Guerra del Ejército del Perú, 2(3), 48-55. http://revistas.esge.edu.pe/RESGE/article/view/30

[13] Poma, A., & Vargas, R. (2019). Problemática en ciberseguridad como protección de sistemas informáticos y redes sociales en el Perú y en el mundo. Sciéndo, 22(4), 275-282. Enlace: https://revistas.unitru.edu.pe/index.php/SCIENDO/article/view/2692

Modelo predictivo de la potabilidad del agua mediante un árbol de decisión en Inteligencia Artificial

Predictive model of water potability through a decision tree in Artificial Intelligence

Angel Alexis Zevallos Apaza
Universidad Nacional de San Agustín. Arequipa, Perú.
azevallosa@unsa.edu.pe

Sofía Sair Onque Gárate
Universidad Nacional de San Agustín. Arequipa, Perú.
sonque@unsa.edu.pe

Arian Eduardo Javier Canaza Cuadros
Universidad Nacional de San Agustín. Arequipa, Perú.
acanazacua@unsa.edu.pe

Paulina Miriam Choqueneira Ccasa
Universidad Nacional de San Agustín. Arequipa, Perú.
pchoqueneira@unsa.edu.pe

ARK: ark:/42411/s9/a72
PURL: 42411/s9/a72

RECIBIDO 28/06/2022 • ACEPTADO 02/08/2022 • PUBLICADO 30/09/2022

RESUMEN

En este trabajo se planteó como objetivo utilizar la técnica de árbol de decisión para definir un modelo capaz de predecir la potabilidad del agua. Para evaluar el rendimiento de la clasificación del árbol de decisión se utilizó un dataset extraído de Kaggle que cuenta con 3276 muestras de agua divididas por la variable de potabilidad. Aplicando las librerías Pandas y Scikit Learn se logró definir un modelo basado en un árbol de decisión evaluado con las métricas de precisión, exactitud, exhaustividad y puntuación F1 logrando 0.77, 0.80, 0.85 y 0.81 respectivamente.

Palabras claves: Agua potable, inteligencia artificial, árbol de decisión.

ABSTRACT

The objective of this work was to use the decision tree technique to define a model capable of predicting water potability. To evaluate the performance of the decision tree classification, a dataset extracted from Kaggle was used, which has 3276 water samples divided by the potability variable. Applying the Pandas and Scikit Learn libraries, a model based on a decision tree

evaluated with the metrics of precision, accuracy, completeness and F1 score was defined, achieving 0.77, 0.80, 0.85 and 0.81 respectively.

Keywords: *Drinking water, artificial intelligence, decision tree.*

INTRODUCCIÓN

La vida del hombre y su existencia se la debe al agua, existe una alta necesidad de esta y cada vez más por el incremento de la población, por consecuencia ocurre una mayor demanda de agua. Existen desigualdades entre las zonas urbanas y rurales, puesto que el 96% de la población mundial urbana utiliza fuentes de agua potable frente al 84% de la población rural, tal como lo dicen en [1], por lo que el poseer agua potable es una necesidad primaria, como lo menciona la Asamblea General de las Naciones Unidas [2].

"Todas las personas tienen derecho a disponer de forma continuada de agua suficiente, salubre, físicamente accesible, asequible y de una potabilidad del agua aceptable, para uso personal y doméstico". Entonces podemos reafirmar que el consumo de agua se debe garantizar, el agua no debe estar contaminada o con sustancias que puedan producir enfermedades ya que este sería un problema muy grave.

Entrando en un contexto cercano, en el Perú más de un 70% de las aguas residuales no tienen tratamiento, la contaminación en el agua puede ser una gran preocupación para nosotros, porque pone en peligro la salud pública. Los principales lugares que superan los límites recomendados por la OMS para el consumo humano de agua son Lima, La Oroya y Juliaca de la cual puede ver más información en [3].

Incluso aquí en Arequipa mediante estudios se determinó que se superaron los parámetros establecidos en bacterias coliformes, que se evalúan en [4]. Por lo que para evitar que esta situación empeore se necesita un buen control del agua verificando que esta sea apta para consumo humano.

"La forma como se mide la contaminación química, los límites que se toleran y las decisiones que se toman al respecto de las afluentes de agua, depende de procesos de monitoreo y vigilancia." [5]. Con esta perspectiva, este trabajo plantea la implementación de inteligencia artificial para determinar las condiciones de la potabilidad del agua a partir de datos que han sido obtenidos de un trabajo de análisis.

Con el modelo se espera determinar con un error mínimo la potabilidad del agua, para beneficiar a la población que obtiene el líquido de afluentes, además de aportar a los procesos que determinan su potabilidad. Ello se plantea hallar en función a los distintos parámetros que influyen

en el resultado de la potabilidad [6], los cuales incluyen el pH, dureza, sólidos disueltos totales, cloraminas, sulfato, conductividad, carbono orgánico, trihalometanos y turbidez.

Materiales y métodos o Metodología computacional

Trabajos relacionados

En [7] se analizó el problema de la contaminación del agua un tema tratado también por nosotros pero con un análisis en tiempo real lo cual tuvo resultados positivos, ya que mediante esta detección se pudo prevenir seguir usando agua contaminada, lo que demuestra que el modelo utilizado en este artículo puede ser muy factible y aplicarse a cualquier variable física que pueda ser medida con un elemento sensor y requiere cierto monitoreo.

En [8] se propuso una solución para determinar el índice de potabilidad del agua del río Utcubamba utilizando redes neuronales artificiales, la RNA planteada fue entrenada usando el algoritmo de Levenberg-Marquardt determinando la distribución óptima consistente en seis neuronas en la capa de entrada, doce neuronas en la capa oculta y una neurona en la capa de salida. La evaluación de la RNA se realizó mediante el coeficiente de correlación y el error de raíz cuadrada media. Los resultados para el coeficiente de correlación para los tres conjuntos de da tos, entrenamiento, validación y prueba fueron 0.979, 1 y 0.940 respectivamente, para el error de raíz cuadrada media para los tres conjuntos de datos fueron 2.562 para entrenamiento, 1.546 para validación y 1.997 para la prueba.

En [9] se mantuvo el problema debido a la población, ya que a más población estas requerirán más necesidades con respecto al agua, pero controla las condiciones apropiadas para asegurar la potabilidad del agua. En muchas situaciones no es suficiente las actividades de monitorización que se brindan, por lo que se requiere contar con modelos o mecanismos que permitan anticiparse a la materialización del riesgo con el suficiente rango de tiempo para prevenir los efectos negativos que afecten la calidad del recurso hídrico. Los resultados obtenidos al final demostraron que la utilización de diferentes técnicas aplicadas, permiten obtener mejores resultados respecto a las técnicas que son utilizadas de forma independiente.

En [10] se propone la revisión de las técnicas de aprendizaje automático y su aplicación para la estimación de la potabilidad del agua en ríos, cuencas y lagos, entre otros, debido a su importancia para la sobrevivencia de los seres vivos. Tras el desarrollo del trabajo, se evidenciaron en los resultados que, pese a que se puede encontrar una gran variedad de estrategias, las redes neuronales han abarcado este campo con buenos resultados, pero aún concentra su atención en los desafíos de combinar sus propiedades para modelar sistemas con características no lineales y no estacionarias. Finalmente, como conclusiones se afirma que las técnicas de aprendizaje automático de mayor aplicabilidad en el recurso hídrico son las redes

neuronales, las máquinas de vectores de soporte y los sistemas de inferencia neuro difusa con porcentajes del 36 %, 24 % y 16 %, respectivamente; el 24% restante corresponde a la implementación de otro tipo de estrategias. Con ello se evidencia que el modelado híbrido es una herramienta mejorada que arroja buenos resultados en comparación con técnicas predictivas tradicionales.

Fundamentación teórica

La inteligencia artificial

La Inteligencia Artificial (IA) es la combinación de algoritmos planteados con el fin de crear máquinas que poseen las mismas capacidades que el ser humano, una ansiada tecnología que todavía resulta lejana y misteriosa, pero que hace algunos años está presente en nuestra vida cotidiana.

Stuart Russell y Peter Norvig diferenciaban la inteligencia artificial en varios tipos, de forma que se tienen los sistemas que piensan como humanos, los que actúan como humanos, los que piensan racionalmente y los que actúan racionalmente.

Los campos de aplicación de la inteligencia artificial ciertamente son muchos y variados, por ejemplo, algunos de los principales son el uso de la IA para los asistentes virtuales, en el campo de la climatología, las finanzas, la agricultura, la educación, la logística y el sistema de transporte, el comercio y los sistemas de sanidad [11].

Machine learning

Machine Learning y el Procesamiento del Lenguaje Natural son campos que convierten a los datos en la información necesaria para alcanzar la Inteligencia Artificial requerida para la extracción de conclusiones, realización de predicciones o comunicación con los usuarios. Las técnicas de Machine Learning permiten a los algoritmos identificar patrones complejos entre grandes cantidades de datos, infiriendo así sus reglas para detectar patrones similares en nuevos conjuntos de datos. Cuando se crean sistemas inteligentes que mejoran de forma autónoma viendo datos, se permite en sí la creación de sistemas que pueden aprender a predecir comportamientos mediante ejemplos, detectar similitudes o anomalías automáticamente o tomar las decisiones adecuadas [12].
Por ellos se dice que mientras dispongamos de más datos, más fiable y representativa será la muestra del proceso que buscamos automatizar, y mejor va a funcionar el software que buscamos generar.

Análisis de datos

El análisis de datos es el proceso de limpieza, cambio y procesamiento de datos en estado bruto, y la extracción de información relevante y procesable que ayuda a tomar decisiones informadas en base a estos. El preprocesamiento ayuda a reducir los riesgos inherentes a la toma de decisiones al proporcionar información y estadísticas útiles, que a menudo se presentan en cuadros o tablas [13].

Árbol de decisión

Un árbol de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas. Este mapa permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. También puede ser usado para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.

Un árbol de decisión, por lo general, comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades. Esto le da una forma similar a la de un árbol.

Algunas de las ventajas de este concepto son su fácil uso y comprensión, la mínima preparación necesaria, la posibilidad de agregar nuevas opciones a árboles ya existentes, así como la facilidad de combinación con otras herramientas de decisión. Sin embargo, también maneja ciertas desventajas, como el hecho de que el árbol se vuelva repentinamente complejo y que demande necesariamente la propuesta de una nueva solución [14].

Ganancia de información

Durante la construcción de un árbol de decisión iremos haciendo varias divisiones y la ganancia de información vendrá a ser precisamente esa información que puede aumentar el nivel de certeza después de una división. Es la entropía de un árbol antes de la división menos la entropía ponderada después de la división por un atributo, por lo que podemos pensar en la ganancia de la información y en la entropía como opuestos [15].

Preprocesamiento de datos

Para comenzar la elaboración de un árbol de decisión, en primera instancia, requerimos de un conjunto de datos (dataset) para trabajar en base a ello. Esta extracción comprende un conjunto de pasos que buscan preparar los datos sin limitarse con la integración u homogeneización, sino

que abarca también otras tareas más, bajo la denominación genérica de preprocesamiento de datos [16].

Una vez seleccionado el dataset, se procede con la limpieza de datos o data cleaning, pues antes de la aplicación de una técnica para procesar y/o analizar un conjunto de datos determinado, es necesario aplicar una limpieza de datos mediante técnicas que permitan filtrar el dataset de los datos vacíos, nulos o incongruentes.

Seguidamente, se tiene el proceso de selección o extracción de datos relevantes, debido a que dos de las tareas más usuales en esta fase son la selección de variables y la selección de patrones. Puntualmente consisten en eliminar aquellos datos que, por estar repetidos o pueden estimarse a partir de otros, no aportan mejora en la extracción de conocimiento.

Finalmente, se tiene la transformación de los datos, teniendo lugar una vez que los datos ya se encuentran limpios y no contienen redundancias, aspectos de los que se ocupan las operaciones previas. En este punto, podría pensarse que ya pueden usarse para el aprendizaje de un modelo, sin embargo, hay acciones que podrían mejorar los datos de modo que haga más efectivo ese aprendizaje. Entre ellos están la normalización, el escalado y la discretización.

Herramientas y elementos

Herramientas

Las herramientas que usamos para el desarrollo del modelo predictivo para asegurar la potabilidad del agua fueron en su mayoría colaborativas para que podamos trabajar en equipo a distancia como Google Colab, el cual nos permite escribir y ejecutar código arbitrario de python en el navegador. Lo tomamos como adecuado para realizar este tipo de tareas y análisis de datos. Para el almacenamiento de nuestra dataSet y los archivos, ya que Google Colab puede trabajar con los documentos de Google, usamos Google Drive, que nos sirvió como sitio de alojamiento de archivos el cual también nos ayudó para trabajar de forma colaborativa. Y para acabar Google Sheet, ya que es ahí en conjunto con Google Drive donde guardamos la dataSet, porque maneja un sistema de celdas, lo cual es permitido para poder arrastrar los datos con python.

Librerías

Usamos librerías para facilitar el desarrollo de nuestro sistema predictivo. Una de ellas fue Numpy, la cual tiene soporte para vector y matrices grandes multidimensionales junto con funciones matemáticas de alto nivel. Para complementar esto y añadirle gráficas a nuestro proyecto, que en nuestro caso fueron árboles de decisión, usamos la biblioteca Matplotib la cual nos sirve para

graficar a partir de datos contenidos en listas o arrays una extensión de esta es NumPy que lo mencionamos antes, también usamos la librería Pandas ya que al nosotros usar dataset necesitábamos tener un mejor control de esta estructura de datos y esta librería es especialista en el manejo y análisis de estructura de datos, como ejemplo de las funcionalidades que nos ofrece son que puede definir nuevas estructuras de datos basadas en los arrays de la librería NumPy pero con nuevas funcionalidades, leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL, acceder a los datos mediante índices o nombres para filas y columnas realizando todas estas operaciones y otras de manera muy eficiente, y como último recurso usamos Scikit Learn una librería para aprendizaje automático la cual suma más ya que esta posee herramientas eficientes usadas para el aprendizaje automático y modelado estadístico incluyendo clasificación, regresión, agrupación, y reducción de dimensionalidad.

Sobre nuestro dataset

Nuestra dataset proviene de [17] en la cual analizamos la potabilidad del agua, teniendo nueve variables de entrada y tres mil doscientos setenta y siete instancias, nuestras entradas para poder obtener un correcto análisis del agua son el ph que posee de 0 a 14, hardness que es la capacidad del agua para precipitar jabón (numérico), solids que nos indica los sólidos disueltos totales (numérico), chloramines que nos dice la cantidad de cloraminas que posee en ppn (numérico), sulfate que nos muestra la cantidad de sulfatos disueltos (numéricos), conductivity que nos dice la cantidad eléctrica del agua (numérico), organic carbon que nos muestra la cantidad de carbono orgánico en ppm (numérico), trihalomethanes que nos dice la cantidad de trihalometanos en ug (numérico), y por último a turbidity que nos dice la medida de la propiedad de emisión de luz de agua en NTU (numérico). La variable a predecir o variable de salida contempla 0 para indicar agua potable y 1 para agua no potable.

Variable	Descripción	Tipo	Min	Max
Ph	Ph del agua	Numérico	1.43	14.00
Hardness	Capacidad del agua para precipitar jabón	Numérico	73.49	306.63
Solids	Sólidos disueltos totales en ppm.	Numérico	320.94	56351.39
Chloramines	Cantidad de cloraminas en ppm.	Numérico	1.39	13.13

Sulfate	Cantidad de sulfatos disueltos en mg/L	Numérico	182.39	481.03
Conductivity	Conductividad eléctrica del agua µS/cm	Numérico	201.62	753.34
Organic carbon	Cantidad de carbono orgánico en ppm	Numérico	2.20	27.01
Trihalomethanes	Cantidad de trihalometanos en µg/L	Numérico	14.34	124.00
Turbidity	Medida de la propiedad de emisión de luz del agua en NTU	Numérico	1.45	6.49
Potability	Indica si el agua es segura para el consumo humano	Numérico	0	1

Tabla N°1 Variables del dataset

4. Resultados y discusión

Los datos que sean utilizado para diseñar el modelo se extrajeron de un dataset que cuenta con 3276 muestras donde se encontró que varias muestras contaban con valores vacíos, las variables Ph, Sulfate y Trihalomethanes contaban con 491, 781 y 162 valores vacíos respectivamente. Para realizar data cleaning en el dataset se optó por la eliminación de datos debido a la cantidad de instancias que se iban a eliminar que alcanzó el 23% del total que no era una cantidad que afecta la consistencia del dataset, luego del proceso se lograron eliminar 1265 instancias vacías.

Se realizó un análisis de las instancias resueltas luego del proceso de data cleaning y se encontró un desbalanceamiento de los datos respecto a la variable a predecir (Potability), contando con 1200 instancias con valor 0 (Potable) y 811 instancias con valor 1 (No potable). Para resolver el desbalance se aplicó la técnica de over sampling definiendo así 2400 instancias en total.

Se empleó un árbol de decisión para determinar si el agua es potable, los datos se dividieron en dos conjuntos, el de entrenamiento y el de pruebas contando con el 80% y 20% de los datos

respectivamente. Luego del entrenamiento y las pruebas se realizó una evaluación del modelo mediante la matriz confusión contando con las métricas de precisión, exactitud, exhaustividad y puntuación F1.

El resultado de la métrica de precisión es de 0.77, refiriendo que, de 100 muestras de agua, el modelo logra reconocer 77 muestras como agua potable de forma correcta mientras que el resto de 23 muestras las clasifica como agua potable de forma incorrecta. En la métrica de exactitud el modelo obtuvo 0.80, es decir, el modelo logra predecir el 80% de las veces, ejemplo, de 100 muestras de agua el modelo clasifica el 80% de forma correcta.

La métrica de exhaustividad del modelo fue evaluada con 0.85, mostrando que de 100 muestras de agua potable el modelo puede reconocer de manera correcta 85 muestras, mientras que el resto se clasifican como agua no potable. Respecto a la puntuación F1 el modelo recibió un puntaje de 0.81 indicando que se reconocen el 81% de los casos positivos donde el agua es potable. El árbol resultado del modelo se muestra en la Imagen N°1.

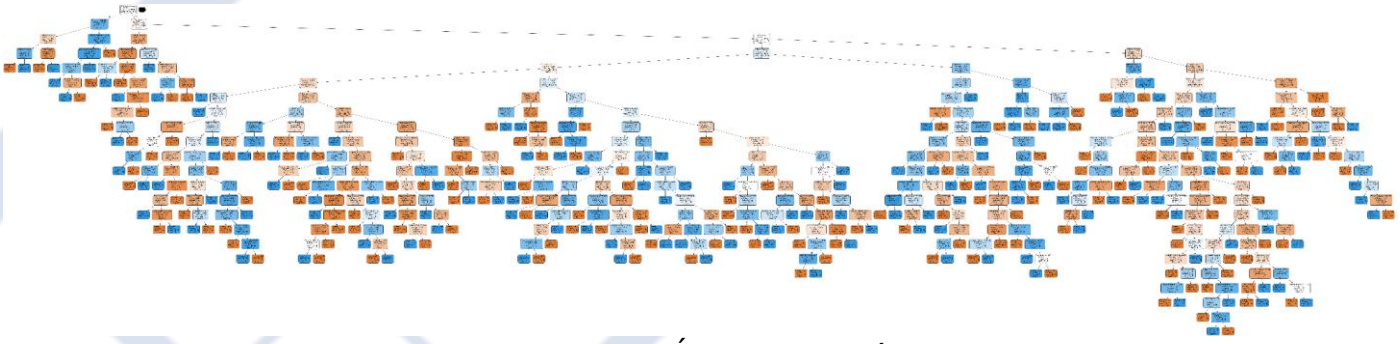

Imagen N°1 Árbol de Decisión

Conclusiones

Se logró definir un modelo capaz de predecir la potabilidad del agua, un modelo diferente al tratado en los trabajos relacionados utilizando una técnica de clasificación. El modelo fue basado en un árbol de decisión que trabajó con un dataset inicialmente con valores vacíos y desbalanceado, el modelo fue evaluado con las métricas de precisión, exactitud, exhaustividad y puntuación F1 obteniendo un puntaje de 0.77, 0.80, 0.85 y 0.81 respectivamente.

El resultado obtenido no logra alcanzar el estándar de 0.80 en todas las métricas, por ello no se puede considerar un gran modelo para la predicción de la potabilidad del agua, sin embargo, se puede concluir que la técnica de árbol de decisión puede ser vista como un poderoso predictor que puede ayudar con la problemática de la potabilidad del agua.

Referencias

[1] C. C. Sánchez, "Enfermedades infecciosas relacionadas con el agua en el Perú," Revista peruana de medicina experimental y salud publica, *35*, 309-316.2018

[2] B. Serrano Pérez, R. Tendero Caballero, & M. D. Río Merino, "Parámetros indicadores del agua potable doméstica urbana, umbrales y consecuencias para la salud," 2018.

[3] F. L. Meoño, C. G. Taranco, & Y. M. Olivares, "Las aguas residuales y sus consecuencias en el Perú. Saber y hacer," 2(2), 8-25. 2015

[4] M. F. Amado Camargo, "Determinación bacteriológica de la calidad del agua de consumo humano, regadío y bebida de animales del Distrito de Majes, Provincia de Caylloma, Departamento de Arequipa, Abril-Mayo 2017", DSpace. Tesis. Arequipa, 2018. Disponible: http://repositorio.unsa.edu.pe/handle/UNSA/5890

[5] A. J. Espinosa Ramírez, "El agua, un reto para la salud pública: la calidad del agua y las oportunidades para la vigilancia en salud ambiental." UNAL. Tesis. Bogotá, 2018. Disponible: https://repositorio.unal.edu.co/handle/unal/63149

[6] C. Idrovo. "Optimización de la planta de tratamiento de Uchupucún," B.S. Tesis. Cuenca, 2010. Disponible: http://dspace.ucuenca.edu.ec/handle/123456789/2426

[7] I. D. López, A. Figueroa y J. C. Corrales, "Un mapeo sistemático sobre predicción de calidad del agua mediante técnicas de inteligencia computacional", Revista Ingenierías Universidad de Medellín, vol. 15, n.º 28, pp. 35–52, 2016. Accedido el 10 de agosto de 2022. Disponible: https://doi.org/10.22395/rium.v15n28a2

[8] L. Quiñones Huatangari, L. Ochoa Toledo, N. Kemper Valverde, O. Gamarra Torres, J. Bazán Correa y J. Delgado Soto, "Red neuronal artificial para estimar un índice de calidad de agua", Enfoque UTE, vol. 11, n.º 2, pp. 109–120, abril de 2020. Accedido el 11 de agosto de 2022. Disponible: https://doi.org/10.29019/enfoque.v11n2.633

[9] A. F. Siles, "Desarrollo de software y diseño de un sistema automatizado para monitoreo y predicción de eventos de contaminación en sistemas de distribución de agua, utilizando inteligencia artificial". Repositorio DspaceDesarrollo de software y diseño de un sistema automatizado para monitoreo y predicción de eventos de contaminación, 1 octubre, 2019.

Accedido el 16 de agosto de 2022. Disponible: http://literatura.ciidiroaxaca.ipn.mx:8080/xmlui/handle/LITER_CIIDIROAX/230

[10] A. C. Aguilar Aguilar y F. F. Obando - Díaz, "APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE CALIDAD DE AGUA POTABLE", Ingeniare, n.º 28, junio de 2020.Disponible: https://doi.org/10.18041/1909-2458/ingeniare.28.6215

[11] Iberdrola. "¿Qué es la Inteligencia Artificial? - Iberdrola". Iberdrola. https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial.

[12] INSTITUTO DE INGENIERÍA DEL CONOCIMIENTO. "Machine Learning y Deep Learning - Expertos en IIC". 2022. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/inteligencia-artificial/machine-learning-deep-learning/.

[13] K. Kelley. "What is Data Analysis? Types, Methods and Techniques 2022, Simplilearn". Simplilearn.com. https://www.simplilearn.com/data-analysis-methods-process-types-article#what_is_data_analysis.

[14] Lucid Software Inc.. "Qué es un diagrama de árbol de decisión". Lucidchart. https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision.

[15] I. Moreno Hojas y StatPlans. "Construyendo árboles de decisión - StatDeveloper". StatDeveloper. https://www.statdeveloper.com/construyendo-arboles-de-decision/.

[16] F. Charte. "Cómo es el proceso de extraer conocimiento a partir de bases de datos - campusMVP.es". campusMVP.es. https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx.

[17] "Water Quality". Kaggle: Your Machine Learning and Data Science Community. 2021, [En línea] https://www.kaggle.com/datasets/adityakadiwal/water-potability.

Gestión de riesgos para el desarrollo de proyectos de sistemas críticos

Risk management for the development of critical systems projects

Guillermo José Aleman Zambrano
Universidad La Salle. Arequipa, Perú.
galemanz@ulasalle.edu.pe
https://orcid.org/0000-0001-5471-4226

Marvik Irzovic Del Carpio Lazo
Universidad La Salle. Arequipa, Perú.
mdelcarpiol@ulasalle.edu.pe
https://orcid.org/0000-0002-0019-2458

Daniel Gustavo Mendiguri Chávez
Universidad La Salle. Arequipa, Perú.
dmendiguric@ulasalle.edu.pe
https://orcid.org/0000-0002-0588-6520

Daniela Carolina Vílchez Silva
Universidad La Salle. Arequipa, Perú.
dvilchezs@ulasalle.edu.pe
https://orcid.org/0000-0002-7896-8228

ARK: ark:/42411/s9/a73
PURL: 42411/s9/a73

RECIBIDO 18/07/2022 • ACEPTADO 02/09/2022 • PUBLICADO 30/09/2022

RESUMEN

Hoy en día podemos encontrar distintos sistemas críticos en diferentes campos como en la salud, militar, espacial, seguridad, etc., dónde peligra la vida y economía de muchas personas debido a las consecuencias que pueden surgir de alguna falla en estos sistemas. Por ello es importante identificar, analizar y tratar los riesgos relacionados a los proyectos de sistemas críticos, procesos típicos de la gestión de riesgos. En este artículo mostramos y analizamos distintas técnicas y modelos aplicados en diferentes ámbitos como la medicina y el campo militar, reconociendo conceptos y similitudes sobre los sistemas críticos y la gestión de riesgos. Finalmente determinando que no existen metodologías definidas para la gestión de riesgos en estos sistemas, siendo en muchos casos necesaria la aplicación de opciones hibridas y dinámicas.

Palabras claves: Gestión de riesgos, Riesgos, Sistemas críticos, Software.

ABSTRACT

Today we can find different critical systems in different fields such as health, military, space, security, etc., where the lives and economy of many people are in danger due to the consequences that may arise from a failure in these systems. For this reason, it is important to identify, analyze and treat the risks related to critical systems projects, typical processes of risk management. In this article we show and analyze different techniques and models applied in different environments such as medicine and the military field, recognizing concepts and similarities about critical systems and risk management. Finally determining that there are no defined methodologies for risk management in these systems, being in many cases necessary to apply hybrid and dynamic options.

Keywords: *Critical Systems, Risks, Risk Managment, Software.*

INTRODUCCIÓN

Los sistemas críticos son aquellos en los que un error podría llegar a significar pérdidas económicas de gran tamaño, daños al medio ambiente, ocasionar daños físicos o amenazar la vida. Muchos sistemas de información modernos están llegando a ser críticos para la seguridad en un sentido general porque pueden producirse pérdidas financieras e incluso la pérdida de vidas debido a su falla. Los futuros sistemas críticos para la seguridad serán más comunes y más poderosos [1]. En pocas palabras para entender que es un sistema crítico debemos enfocarnos en sus consecuencias, si la falla de un sistema puede tener consecuencias que se consideran inaceptables, entonces el sistema es crítico para la seguridad.

En esencia, un sistema es crítico para la seguridad cuando dependemos de él para nuestro bienestar [1]. También se considera que algunos subsistemas que conforman el sistema crítico pueden llegar a ser considerados sistemas críticos dependiendo de las consecuencias provocadas por su falla. Estos sistemas los podemos encontrar en el campo militar [2], médico [3], espacial [4], seguridad [5] y entre otros [1], al contar con consecuencias graves es importante aplicar una gestión de riesgos a este tipo de proyectos, para así determinar el alcance de los daños y los controles que se aplicarán para disminuir o eliminarlos.

Para poder mitigar los riesgos en diferentes sistemas, se aplica un concepto de sistema superviviente, el cual consiste en asegurar la confiabilidad del software a través de una alta disponibilidad de sus servicios, de forma tal que sean siempre por lo menos óptimos. Así un sistema crítico superviviente puede dar la seguridad de continuar incluso en caso de fallas, dentro del sistema global. Esto se logra dando prioridad a los subsistemas que hacen critico a todo el sistema, sacrificando al resto de subsistemas cuando sea necesario [7]. Así hacer que un sistema crítico puede aplicarse a los distintos ámbitos de clasificación de sistemas críticos.

Los pueden ser clasificados por Tradicionales y no tradicionales. Los primeros sistemas críticos incluyen la atención médica, los aviones comerciales, la energía nuclear y los armas. La falla en estas áreas puede conducir rápidamente a que la vida humana se ponga en peligro, la pérdida de equipos, etc. [1] Los segundos son aquellos sistemas que a simple vista no son sistemas críticos, pero podrían llegar a serlo si su falla se prolonga o sus consecuencias escalan [1]. Un ejemplo que [1] muestra es la pérdida del sistema telefónico que a simple vista no puede causar la muerte de personas. Pero una pérdida prolongada del servicio 911 sin duda resultará en lesiones graves o la muerte. El servicio de emergencia 911 es un ejemplo de una aplicación de infraestructura crítica. Otros ejemplos son el control del transporte, los sistemas bancarios y financieros, generación y distribución, telecomunicaciones y la gestión de los sistemas de agua.

Una mala planificación de la gestión de riesgos dentro de la elaboración del proyecto implicaría fallas que, si se llegan a lanzar a producción, llegaría así a ser contraproducente para la empresa y en caso de no llegarse a culminar el proyecto implicaría grandes pérdidas por la inversión que implican el desarrollo de proyectos orientados a sistemas críticos. Existen varios ejemplos de fallas en sistemas críticos que han ocurrido, incluidos la falla de la cuenta regresiva del transbordador espacial en el primer lanzamiento, la falla del lanzamiento del Ariane V/5 [8] y las pérdidas del Mars Polar Lander [9] y el Mars Climate Orbiter [10]. Podemos encontrar otros ejemplos documentados en el texto de Neumann [11] que analiza una gran colección de problemas experimentados a lo largo del tiempo y proporciona ideas que pueden ser útiles para evitar tales consecuencias en el futuro.

Para el PMBOK [6] del Project Management Institute el riesgo es un evento o condición incierta que sí ocurre, tendrá un efecto positivo o negativo en los objetivos del proyecto. De aquí la necesidad de su gestión para promover las ventajas o mitigar las desventajas que estén relacionadas a cada riesgo que se puede identificar. Una buena gestión de los mismos asegura que los recursos empleados en el proyecto tengan un alto rendimiento sobre el logro de los objetivos planteados para el desarrollo.

Entonces al existir diversidad de riesgos a identificar, su clasificación es parte importante para una correcta gestión, es así que este trabajo evalúa las técnicas aplicadas a la gestión de riesgos en distintos campos donde se desarrolla proyectos orientados a sistemas críticos.

Materiales y métodos o Metodología computacional

Esta investigación se basó en una metodología descriptiva, basándonos en los sistemas críticos y sus aplicaciones como objeto de estudio mismo. El documento se estructura en dos secciones base para finalizar en la conclusión. La primera parte consiste en la recopilación de información sobre sistemas críticos en diversos ámbitos como la medicina, las aplicaciones militares, los usos

en el transporte, y sistemas ciber físicos en general que permiten el control del entorno real a través de la administración de mecanismos controlados por software. La segunda parte consiste en la abstracción de los conocimientos adquiridos y la comparación de los mismos para mostrar los resultados obtenidos. Logrando esto a través del consenso de ideas entre los distintos autores, y la identificación de sus diferencias clave, que contribuyen al enriquecimiento de ideas sobre el tema.

Finalmente se explican las conclusiones consecuencia de la información y abstracción del conocimiento de las dos primeras partes. Comparando previamente resultados y propuestas de las aplicaciones de los sistemas críticos estudiados.

Resultados y discusión

Luego de la realizar una ardua investigación pudimos observar que en el campo de la medicina encontramos un ejemplo de proyecto [12] que tiene como objetivo analizar la aplicación de la metodología FMEA (Análisis de Modos de Falla y Efectos) y sus efectos de mejora en un sistema médico para distribución de medicamentos.

> *FMEA es un proceso de mejora de la calidad que se concentra en el sistema general en un entorno en lugar de asignar todos los errores al error humano. Esta técnica divide un proceso determinado, como un sistema de distribución y administración de medicamentos, en pequeños pasos. Los métodos de aseguramiento de la calidad utilizados en otras industrias, como sacrificar un porcentaje de un lote de producto para la prueba, no se pueden aplicar en situaciones de administración de medicamentos que involucran a seres humanos. El objetivo de FMEA es utilizar la experiencia de las personas en el campo para evaluar un sistema y anticipar las posibles formas en que puede fallar. Una vez que se han establecido los modos y mecanismos de falla básicos, se clasifica la importancia relativa de cada uno para el sistema general [12].*

El caso de estudio se realizó en el Hospital Sir Charles Gairdner [12] y para aplicar la técnica se instauró un observador farmaceuta, quien era encargado de registrar todos los hallazgos de las áreas del proceso involucradas en el sistema de suministro de medicamentos. con estos hallazgos se detectaron los riesgos y fallas junto a su área perteneciente. Además, gracias a la técnica se pudo determinar la tasa y los tipos de errores de medicación que ocurrían en el sistema de existencias del hospital. Las posibles fallas se identificaron prediciendo qué acciones incorrectas podría realizar una persona, cuáles podrían ser los resultados de esas acciones incorrectas y cómo se podrían prevenir las acciones incorrectas, gracias esto se identificó las deficiencias del sistema y desencadenó a desarrollar un nuevo sistema. En cuanto a los resultados "*El análisis del sistema*

de administración y distribución de medicamentos identificó 12 fallas del sistema. Se eliminaron o redujeron una serie de posibles errores de medicación al pasar del sistema de existencias de la sala al nuevo sistema. Las limitaciones del estudio incluyen el hecho de que menos del 100% de las enfermeras se ofrecieron como voluntarias para ser observadas durante el estudio" [12].

Se observa que la técnica aplicada demanda la atención e intervención de una persona, ya que al ser parte de un sistema crítico no podemos aplicar técnicas débiles para el análisis e identificación de riesgos, porque cualquier error desencadenaría consecuencias graves. Para el caso de estudio los resultados de la aplicación de FMEA obtuvieron nuevos requisitos y mejoras que lo llevaron a desarrollar un nuevo y mejorado sistema.

Otra aplicación con respecto al campo de medicina en la gestión de riesgos, fue encontrada en el campo de Internet de las Cosas Médicas (IoMT), con una interconexión de dispositivos médicos [13]. Debido a la naturaleza del campo y a que los dispositivos IoMT se conectan a sistemas para su comunicación e integración, esto los vuelve parte de sistemas críticos. Además, estos sistemas manejan una gran data y están estrechamente relacionados con pacientes por lo que desencadenaría consecuencias desastrosas. Se utiliza una matriz de Severidad-Probabilidad que permite detectar los fallos gráficamente, cruzando ambos criterios para determinar si el riesgo es aceptable, razonablemente práctico-tolerable, e intolerable [13].
Demostrando que aplicar estos criterios mejora la confiabilidad de sus sistemas, ya que esta demanda un alto nivel de seguridad, al estar inmersa en las redes es vulnerable a una gran cantidad de ataques. Claramente se ve la necesidad e importancia de una gestión de riesgos en este campo de la tecnología.

En el campo militar [2] encontramos un ejemplo de proyecto que tiene como objetivo el poder reconocer todos los posibles riesgos que podrían presentarse en el software de aeronaves militares y en este caso utilizaron Software Hazard Analysis (SWHA) esta metodología se centra en la identificación de los riesgos asociados con el diseño en el que se evaluarán restricciones operativas para mejorar aún más los requisitos de diseño y los esfuerzos de prueba para el software crítico para la seguridad, en conclusión el SWHA es esencialmente un análisis de requisitos de seguridad, el SWHA utiliza un índice de riesgo de software el cual sirve como una guía para la ingeniería de seguridad, el proceso de desarrollo e integridad y la gestión de programa que otorgan la cantidad adecuada de esfuerzo para garantizar la seguridad del sistema, también utiliza la evaluación de riesgo, esta nos indica el nivel de gravedad del percance dentro del contexto del sistema y de la organización, para ello se utiliza la tabla de condición de falla donde se categoriza la gravedad de los accidentes por catastróficas, criticas, marginales y negligentes, la siguiente clasificación es la control de software esta nos muestra la condición en la que se encuentra por cuatro niveles, el primer nivel nos indica que la falla ante la prevención de un evento nos podría conducir directamente a un peligro, el nivel dos se divide por en dos, primero nos dice que se debe dar tempo para la intervención de un sistema de seguridad independiente para mitigar el peligro para luego tener una acción inmediata para mitigar el

peligro, el nivel tres indica que se necesita de una acción humana para completar la función de control para luego con ayuda del sistema que genera información crítica poder tomar una decisión y por último el nivel cuatro el cual es el que muestra mayor independencia e indica que el software directamente no controla el sistema crítico y no proporciona información que nos ayude a tomar decisiones por ende el sistema es totalmente independiente al riesgo, estas dos tablas son fundamentales para generar la matriz del índice de riesgos de software, gracias a este análisis se pudieron encontrar las fallas y poder generar una correcta clasificación de riesgos.

También pudimos observar que en el artículo [5] se plantean diversas opciones para la gestión de riesgos en sistemas ciber físicos (CPS), estos son sistemas inteligentes que incluyen redes de interacción diseñadas de componentes físicos y computacionales, en otras palabras, en un mecanismo (sistema físico) controlado o monitorizado por algoritmos basados en computación y que a menudo el intercambio de datos lo realizan por medio de internet en tiempo real. Ejemplos de sistemas ciber físicos serían los famosos IOT o Internet de las cosas. Estos sistemas podrían ser considerados como críticos al producir consecuencias muy graves en el bienestar de las personas tanto a nivel económico y salud. Los CPS abarcan distintos dominios los cuales incluyen sistemas biomédicos y de salud, sistemas de transporte, sistemas automotrices, y sistemas de fabricación [5]. Los sistemas ciber físicos se encuentran expuestos a una serie de riesgos que importante controlar. Por ejemplo, al estar integrado a internet podría sufrir ataques cibernéticos, posible robo de información, etc. Para ello el artículo [5] plantea que es importante aplicar diferentes técnicas la para evaluación de riesgos, así como técnicas de reducción. También recomienda que las herramientas e instrumentos desarrollados para la gestión de riesgos deben ser rápidos, rentables y prácticas.

Mencionan a su vez que recibieron diferentes propuestas para esta gestión de riesgos entre las que destacan la de Aakarsh Rao y sus colegas quienes en [14] presentan un enfoque dinámico de gestión y mitigación de riesgos basado en la estimación probabilística de amenazas. Con el objetivo de garantizar la seguridad, la protección y la privacidad en presencia de amenazas de seguridad desconocidas, los dispositivos deben detectar y evaluar el riesgo de forma dinámica y, posteriormente, tomar medidas de mitigación automatizadas cuando el riesgo sea elevado, Aakarsh Rao y su equipo propuso un modelo incorporado un novedoso detector de amenazas en tiempo real con una metodología de evaluación de riesgos adaptativa para garantizar una mitigación de amenazas completa durante la implementación de dispositivos. Para la demostración de su funcionamiento desarrollaron un prototipo de marcapasos con conexión inteligente en la cual implantaron un malware. Los componentes críticos necesarios para el rendimiento esencial del marcapasos incluyen el marcapasos, el sensor y el componente de cómputo de estimulación. El middleware hardware-software facilita la transferencia segura de datos y señales entre modos operativos. El middleware también es responsable de analizar la detección de amenazas en tiempo de ejecución, la actualización del modelo de riesgo y la determinación de qué estrategia de mitigación invocar cuando se detecta una amenaza.

Conclusiones

La gestión de riesgos para los sistemas críticos se basa principalmente en la determinación y valoración de riesgos como consecuencia de una metodología determinada. Justamente esta metodología varía según cada investigación, dando como resultados en todas ellas mejoras de los análisis de riesgos lo que permitió mitigar los riesgos, o al menos mejorar los ámbitos donde se desencadenan para mitigar sus efectos.

Los sistemas críticos están presentes en diferentes campos y aplicaciones, teniendo todas en común la importancia de conservar sus sistemas funcionando constantemente y de manera correcta.

Muchos sistemas no son críticos de forma individual, sin embargo, pueden ser relacionados un sistema crítico, o incluido como parte de uno. En el primer caso el software no critico se vuelve crucial, al ser un soporte critico de un sistema crítico, lo que por transitividad ambos sistemas se vuelven críticos. En el segundo caso el planteamiento es similar, con la única distinción que el software no critico es incluido en el software crítico, siendo un soporte embebido en contraposición al primer caso.

A diferencia del punto anterior también se puede se puede dividir los sistemas en subsistemas, permitiendo identificar los subsistemas críticos y no. Lo que contribuye a una gestión de riesgos más incidente sobre lo realmente crítico, derivando tiempo y recursos a lo más importante. Esto a su vez permite la creación de sistemas supervivientes que siempre estén disponible a un nivel óptimo de servicio, aunque no sea el mejor todo el tiempo.

El análisis de riesgos presenta varias técnicas para su identificación, no se puede determinar una técnica ideal para aplicar a todos los proyectos. Esta técnica dependerá del ámbito, especialidad y contexto de desarrollo del proyecto, incluso aplicará en determinadas ocasiones aplicar técnicas hibridas, o dinámicas para adaptar las metodologías y lograr los mejores resultados.

Referencias

[1] Knight, J. C. (2002, May). *Safety critical systems: challenges and directions. In Proceedings of the 24th international conference on software engineering* (pp. 547-550).

[2] Oh, H. J., & Hong, J. P. (2012). A Study of Software Hazard Analysis for Safety Critical Function in Military Aircraft. *Journal of IKEEE*, *16*(2), 145-152.

[3] Gatouillat, A., Badr, Y., Massot, B., & Sejdić, E. (2018). Internet of medical things: A review of recent contributions dealing with cyber-physical systems in medicine. *IEEE internet of things journal*, *5*(5), 3810-3822.

[4] Albee, A., Battel, S., Brace, R., Burdick, G., Casani, J., Lavell, J., ... & Dipprey, D. (2000). Report on the loss of the Mars Polar Lander and Deep Space 2 missions.

[5] Biro, M., Mashkoor, A., Sametinger, J., & Seker, R. (2017). Software safety and security risk mitigation in cyber-physical systems. *IEEE Software*, *35*(1), 24-29.

[6] Guide, A. (2001). *Project management body of knowledge (pmbok® guide). In Project Management Institute* (Vol. 11, pp. 7-8).

[7] Knight, J. C., & Strunk, E. A. (2004). Achieving critical system survivability through software architectures. In *Architecting Dependable Systems II* (pp. 51-78). Springer, Berlin, Heidelberg.

[8] Lions, J. L., Luebeck, L., Fauquembergue, J. L., Kahn, G., Kubbat, W., Levedag, S., ... & O'Halloran, C. (1996). Ariane 5 flight 501 failure report by the inquiry board.

[9] Albee, A., Battel, S., Brace, R., Burdick, G., Casani, J., Lavell, J., ... & Dipprey, D. (2000). Report on the loss of the Mars Polar Lander and Deep Space 2 missions.

[10] Board, M. I. (1999). Mars Climate Orbiter Mishap Investigation Board Phase I Report November 10, 1999.

[11] Neumann, P. G. (1994). Computer-related risks. Addison-Wesley Professional.

[12] McNally, K. M., Page, M. A., & Sunderland, V. B. (1997). Failure-mode and effects analysis in improving a drug distribution system. American Journal of Health-System Pharmacy, 54(2), 171-177.

[13] Gatouillat, A., Badr, Y., Massot, B., & Sejdić, E. (2018). Internet of medical things: A review of recent contributions dealing with cyber-physical systems in medicine. *IEEE internet of things journal*, *5*(5), 3810-3822.

[14] Rao, A., Carreón, N., Lysecky, R., & Rozenblit, J. (2017). Probabilistic threat detection for risk management in cyber-physical medical systems. IEEE Software, 35(1), 38-43.

Sistema de identificación de emociones a través de reconocimiento facial utilizando inteligencia artificial

Emotion identification system through facial recognition using artificial intelligence

Alexandra Paricela Canazas
Universidad Nacional de San Agustín.
Arequipa, Perú.
aparicelac@unsa.edu.pe

Johnnathan Jimmy Ramos Blaz
Universidad Nacional de San Agustín.
Arequipa, Perú.
jramosb@unsa.edu.pe

Patricio Dante Torres Martínez
Universidad Nacional de San Agustín.
Arequipa, Perú.
ptorresmar@unsa.edu.pe

Xiomara Jaquehua Mamani
Universidad Nacional de San Agustín.
Arequipa, Perú.
xjaquehua@unsa.edu.pe

ARK: ark:/42411/s9/a74
PURL: 42411/s9/a74

RECIBIDO 20/08/2022 • ACEPTADO 24/09/2022 • PUBLICADO 30/09/2022

RESUMEN

El presente articulo tiene como principal objetivo el desarrollo de un sistema que permita identificar las emociones de una persona mediante el reconocimiento de rostros utilizando inteligencia artificial. Para el desarrollo del sistema se tuvo como base el algoritmo básico de Eigenfaces o Análisis de Componente Principal, el cual es uno de los modelos de reconocimiento de rostros más utilizado. Así mismo fue utilizado el lenguaje Python y algunas de sus librerías disponibles como Numpy, OpenCV y Sklearn para la implementación.

Palabras claves: Expresiones faciales; Emociones; Visión por computadora; Aprendizaje automático; Inteligencia artificial.

ABSTRACT

The main objective of this paper is the development of a system to identify the emotions in a person by means of face recognition using artificial intelligence. The system development was based on the basic algorithm of Eigenfaces or Principal Component Analysis, which is one of the

most widely used face recognition models.In addition, Python language and some of its available libraries as Numpy, OpenCV y Sklearn were used for the implementation.

Keywords: *Facial expressions; Emotions; Computer Vision; Machine Learning; IA.*

INTRODUCCIÓN

La influencia de las emociones es crucial en el proceso de aprendizaje, y es, además, inherente al lenguaje y las formas empleadas para la transmisión de la información; por consiguiente, es importante comprender cómo es que las emociones modifican el proceso de aprendizaje y cómo a partir de esto se diseñan modelos que buscan optimizar los logros de aprendizaje. Para esto se consideran cinco factores o estímulos: fisiológicos (auditivo, visual, táctil y cenestésico), sociológicos, ambientales, psicológicos y emocionales que afectan el aprendizaje de un individuo [1]. Así, podemos destacar aspectos como la personalidad, el estado de ánimo y las emociones que juegan un papel importante en los entornos de aprendizaje.

Por otra parte, también es claro que los niños recurren a sus emociones con el fin de dar sentido a su pertenencia al participar en el proceso de aprendizaje; y es este sentido de pertenencia el que define la forma en la que encuentran significativa su participación [2]; algo muy asociado a la idea de libre elección y expresión emocional.

A pesar de esto, dado el caso, puede darse que no exista relación entre las emociones y la ganancia proporcional de aprendizaje; sin embargo, una significativa correlación entre las emociones y la precisión de algunos procesos cognitivos y metacognitivos [3].

Enfocado en la psicología: En los últimos años y con el avance de la tecnología se han dado en mayor parte las consultas a psicólogos o tutores privados por internet u online sin embargo el desempeño psicológico de un estudiante puede llegar a verse afectado por el sobrecargo de tareas o la escuela para notar estos rasgos es necesario un análisis completo de su comunicación no verbal como lo que expresa los gestos que muestra [4]. Para ayudar en este ámbito tanto a los tutores como psicólogos u otro oficio relacionado se plantea una IA que permite entrenar un sistema en el que a través de cámaras se registre los gestos fáciles y a través de esta se identifiquen las emociones de los estudiantes en instituciones educativas [5].

Comúnmente los estudiantes pueden tener problemas familiares, sociales, personales que lo pueden afectar mental y físicamente; teniendo esto en cuenta los profesores/encargados no notarían siempre, o que obvian esto, sin darse cuenta de las consecuencias psicológicas que pueden ocasionar a lo largo de los años. Por ello se elaboraría un sistema que de precaución con conocimientos similares a los que poseen los humanos. [6] ante las expresiones de los estudiantes

que será clasificado con las expresiones faciales prototípicas básicas reconocidas universalmente. Estas son: enojo, disgusto, miedo, felicidad, tristeza y sorpresa. [7].

En este artículo se tratará de abordar un sistema enfocado a base en reconocimiento facial buscando los patrones y características de los rasgos faciales captados dentro de la imagen, para que esta sea recortada solo al contorno de la cara [8], entre ellas se debe hacer una identificación y análisis de cada tipo de emociones con su intensidad para así calcular el porcentaje y la intensidad de cada emoción [9] [10].

Materiales y métodos o Metodología computacional

Estado del Arte

A. Reconocimiento facial con deep learning y Python

El presente trabajo [6] describe e implementa los pasos necesarios para crear un programa capaz de identificar a las personas presentes en una imagen, video o webcam haciendo uso de las librerías disponibles en python y aplicando los conceptos de Deep Learning y redes neuronales.

B. Detección de emociones y reconocimiento facial utilizando aprendizaje profundo

El presente trabajo de titulación [4] tiene como finalidad implementar un sistema de detección de emociones y reconocimiento facial a través del aprendizaje profundo, específicamente utilizando la técnica de redes neuronales convolucionales y la librería face recognition. De este modo se crea una arquitectura de aprendizaje profundo con dataset de Kaggle denominado "Learn facial expressions from an image" con el algoritmo Frontal Face elaborada en Django.

C. Uso de patrones de reconocimiento de las emociones para apoyar la didáctica de enseñanza aprendizaje

El articulo [5] presenta el desarrollo de una investigación aplicada en el ámbito de la teoría educativa apoyado por procesos en comunicación, con el fin de permitir el mejoramiento de las estrategias de enseñanza en el ámbito superior técnico profesional o nivel terciario en educación, considerando el uso de patrones de reconocimiento para capturar las emociones según el registro de reacciones faciales haciendo uso de los algoritmos "Eigenfaces" y "Fisherfaces" y a través de ellos aplicaciones de filtros para regular las distorsiones de la digitalización.

D. Diseño de un Sistema de Reconocimiento de rostros aplicando inteligencia y visión artificial

El presente articulo [9] se centra en el proceso de diseño, desarrollo e implementación de un sistema de reconocimiento de rostros mediante la hibridación de técnicas de reconocimiento de patrones, visión artificial e inteligencia artificial con el uso de redes neuronales. Así mismo es recopilado el producto de la unión de las técnicas de visión artificial y las técnicas de inteligencia artificial y sus implicaciones en múltiples aplicaciones tales como el control de robots de interacción social.

E. Sistema de reconocimiento de gestos faciales captados a través de cámaras para analizar el nivel de satisfacción de clientes en restaurantes

La presente investigación [8] tiene como principal objetivo el desarrollo de un sistema que reconozca la satisfacción o insatisfacción de un cliente en un restaurante con base en los gestos que este mismo realiza al momento de recibir el servicio brindado por el establecimiento. En la etapa de preprocesamiento de los datos se hizo uso de una máquina de soporte vectorial como clasificador, a su vez el histograma de gradientes fue usado en la detección de rostro dentro de la imagen.

Fundamentación Teórica

Reconocimiento Facial

Para el desarrollo del proyecto, centrado en el reconocimiento de emociones, era indispensable la consideración de un sistema de reconocimiento facial, y más específicamente, el algoritmo seleccionado para esto fue el de Eigenfaces, uno de los modelos de reconocimiento de rostros más utilizado, el cual está basado en las propiedades matemáticas de la imagen digitalizada, que captura características invariantes de los rostros.

No fue hasta la década de 1960 que se desarrolló el primer sistema semiautomático de reconocimiento facial basado en un método mediante el cual el observador localizaba rasgos faciales en fotografías, a partir de esto, se buscó el cálculo de distancias y proporciones con el fin de realizar comparaciones.

Ya en la década de 1970, se propuso el uso de características específicas para el reconocimiento facial y posteriormente en 1988, Sirvoich y Kirby propusieron el método de análisis de componentes principales (PCA), dando inicio a la mejora en este ámbito.

Es por esto que actualmente se cuenta con diversas técnicas para el procesamiento de imágenes y el reconocimiento de patrones, convirtiéndose en un área de investigación activa [11].

Reducción de Dimensionalidad

La reducción de dimensionalidad se refiere a técnicas o algoritmos que buscan una proyección de la representación original de muchas dimensiones de nuestros datos a una de menor dimensión, reduciendo el número de variables de entrada en los datos de entrenamiento.

Es decir, buscan una superficie de baja dimensión, incrustada en el espacio de datos de alta dimensión, proyectando los datos a un subespacio dimensional más bajo que capturan lo más importante de los datos, de modo que al tratar menos dimensiones también se trabaje con menos parámetros, simplificando el modelo de aprendizaje automático [12].

Es por esto que la reducción de la dimensionalidad no implica pérdida de información; sin embargo, pueden emplearse representaciones con pérdida y a pesar de esto obtener un buen control, lo que requerirá un equilibrio cuidadoso entre preservar aspectos importantes de las distribuciones y usar la menor cantidad de dimensiones posible.

Análisis de Componentes Principales

Una de las técnicas de reducción de dimensionalidad más comunes es el Análisis de Componentes Principales, en el que, dado un conjunto de datos, se busca la representación lineal de menor dimensión de los datos conservando la varianza de los datos reconstruidos. De esta forma, esta técnica hace uso del hiperplano de baja dimensión tal que, al proyectar los datos en este hiperplano, los cambios en la varianza de los datos sean mínima [13].

Así, aplicando PCA, el vector de primera base, u1, apunta en la dirección en la que los datos tienen la mayor varianza. En otras palabras, las proyecciones de los puntos de datos sobre u1, dadas por $\mathrm{u}\frac{T}{1}\,\mathrm{x1},\ldots,\mathrm{u}\frac{T}{1}\,\mathrm{xN}$, tienen la varianza muestral más grande posible, para cualquier elección del vector unitariou1.El segundo vector base u2, apunta en la dirección de máxima varianza sujeto a la restricción de que u2 sea ortogonal a u1(es decir,$\mathrm{u}\frac{\mathrm{T}}{2}\,\mathrm{u1} = 0$). En general, el i-ésimo vector base, también llamado i-ésimo componente principal, es la dirección de máxima varianza que es ortogonal a los $i - 1$ componentes principales anteriores.

Eigenfaces

En 1991, Turk y Pentland propusieron el enfoque Eigenfaces para el reconocimiento de rostros que utiliza conceptos de álgebra lineal y reducción de dimensionalidad para reconocer rostros, tras los fundamentos planteados por Sirovich y Kirby en 1987 en su artículo Procedimiento de baja dimensión para la caracterización de rostros humanos, convirtiéndose en un enfoque fundamental en la historia de la visión computacional.

En este enfoque, al ingresar un conjunto de datos de Z imágenes de rostros es necesario que cada rostro se represente en escala de grises como un mapa de bits NxN de los píxeles. Es necesario formar un solo vector a partir de la imagen, y esto se consigue simplificando cada una en un vector de características de intensidades de píxeles sin procesar; posteriormente se concatenan todas las filas juntas, formando una única y larga matriz de intensidades de píxeles en escala de grises de Z filas y N2 columnas.

A partir de esa matriz se calcula la media de cada columna en la matriz, obteniendo el valor de intensidad de píxel promedio para cada coordenada en el conjunto de datos de la imagen. Es necesario centrar los datos en la media restando está a cada columna para obtener la matriz de covarianza y calcular los valores propios y los vectores propios de la matriz de covarianza.

A partir de los valores propios más grandes, se calculan los vectores propios correspondientes y se obtiene un lapso dimensional. Haciendo uso de las caras de entrenamiento normalizadas se representa cada vector de rostro en la combinación lineal de los mejores vectores propios [14].

Herramientas y Elementos

Google Colab Es una herramienta para escribir y ejecutar código Python en la nube de Google. También es posible incluir texto enriquecido, links e imágenes. En caso de necesitar altas prestaciones de cómputo, el entorno permite configurar algunas propiedades del equipo sobre el que se ejecuta el código. El uso de Google
Colab permite disponer de un entorno para llevar a cabo tareas que serían difíciles de realizar en un equipo personal. [15].

Google Drive Es un servicio de almacenamiento de datos que son guardados en la nube. Permite copiar archivos desde el ordenador para que sean guardados en la nube [16]. Para el presente trabajo se hará uso del
mismo para almacenar las carpetas de nuestro dataset para el entrenamiento de la IA, así mismo las carpetas
para la validación y resultados al poner a prueba el reconocedor.

Python Es un lenguaje de programación potente y fácil de aprender. Tiene estructuras de datos de alto niveles eficientes y un simple pero efectivo sistema de programación orientado a objetos [17]. Por lo tanto, es ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML).

Las librerías utilizadas de este lenguaje en el presente proyecto fueron las siguientes:

Numpy NumPy es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. Es utilizada para trabajar con matrices permitiendo representar colecciones de datos de un mismo tipo en varias dimensiones, y funciones muy eficientes para su manipulación. NumPy significa Python numérico. [18]

Os El módulo OS en Python proporciona funciones para interactuar con el sistema operativo. OS viene bajo los módulos de utilidad estándar de Python. Este módulo proporciona una forma portátil de usar la funcionalidad dependiente del sistema operativo. [19]

OpenCV-Python OpenCV-Python es una biblioteca de enlaces de Python diseñada para resolver problemas de visión por computadora. Además, es compatible con las plataformas más utilizadas, Windows, Mac OS y Linux [20].

Sklearn Scikit-learn (Sklearn) es la biblioteca más útil y sólida para el aprendizaje automático en Python. Proporciona una selección de herramientas eficientes para el aprendizaje automático y el modelado estadístico, incluida la clasificación, la regresión, el agrupamiento y la reducción de la dimensionalidad a través de una interfaz de consistencia en Python. [21]

Información del Dataset

El dataset que se consideró utilizar para el entrenamiento de la IA se encuentra en kaggle con una cantidad de 28 821 imágenes cada una por defecto separada en carpetas con su emoción distinta, estas imágenes ya cuentan con la escala de grises y la medida sugerida para la detección de rostros.

Para el caso del entrenamiento se hizo el cambio de unir todas las imágenes separadas en carpetas en una sola, y con esta por sí sola sea el entrenamiento para el reconocedor, y que este identifique por si mismo los rasgos considerados comunes y con ello distinguiera la diferencia y similitud entre estos para el reconocimiento de próximas imágenes.

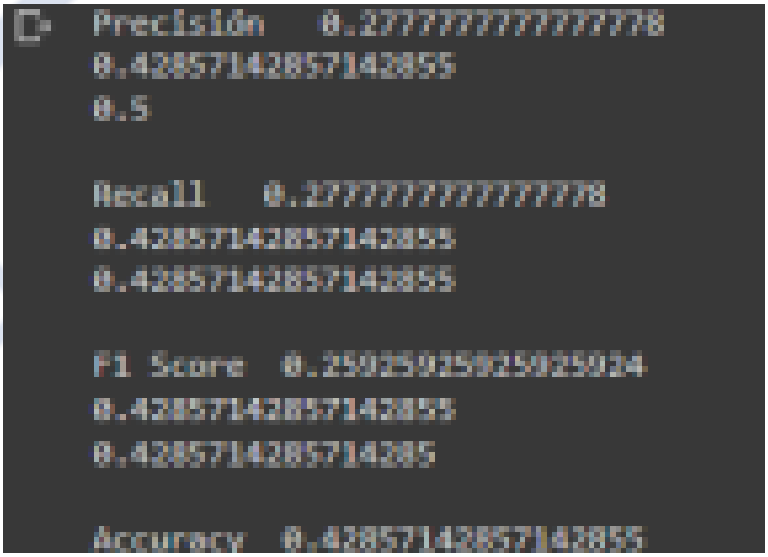

Figura 1. Resultado

Figura 2. Resultado- 2

Resultados y discusión

La calidad del cálculo de la precisión contando el total de verdaderos positivos, falsos negativos y falsos positivos es 0.7142857142857143. Por lo que se interpretaría que un 43 % de los casos en el que se equivocaría el reconocedor.

La cantidad que el reconocedor es capaz de identificar se encuentra el cálculo de forma global Recall= 0.5714285714285714 En el rendimiento promediado de la precisión y la exhaustividad entre varias soluciones del F1 score comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones 0.619047619047619, indicando que el rendimiento del reconocedor.

Como ultimo resultado de la Exactitud (Accuracy) mide el porcentaje de casos que el reconocedor ha acertado, en este caso el resultado es 0.5714285714285714.

Conclusiones

El reconocedor al tener un entrenamiento tiene una mayor identificación de rostros de diferentes medidas y calidades. El lenguaje de programación Python da bastante acceso con las librerías a la elaboración del entrenamiento y del reconocedor de detección de rostros.

Las características faciales tienen mucha similitud al momento de la identificación de emociones, por lo que se no se pueden basar con una sola característica para la identificación correcta de la emoción.

El entrenamiento debe ser específico según el objetivo que se quiera llegar, para cada acción específica que se quiera conseguir, se necesita otro entrenamiento enfocado en ese objetivo. El sistema tuvo en total una exactitud de aciertos ligeramente positiva por el entrenamiento que no fue muy eficiente para el modelo de Eigen Face.

Referencias

[1] J. S. P. Doulik and I. Simonova, Learning Styles in the e-Learning Environment: The Approaches and Research on Longitudinal Changes., 2017.

[2] S. Frankel and M. Mountford, In search of meaningful participation: Making connections between emotions and learning., 2021.

[3] R. R. E. B. C. G. B. M. Taub, R. Azevedo and M. J. Price, How are students^a€™ emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?, 2021.

[4] J. Romero, Detecci´on de Emociones y Reconocimiento Facial utilizando aprendizaje profundo, 2020.

[5] J. L. Mancilla Monsalve, Uso de patrones de reconocimiento de las emociones para apoyar la did´actica de ense˜nanza aprendizaje., 2019.

[6] S. Roy, Face emotion recognition with EfficientNetB2, 2021.

[7] E. Jes´us, Detecci´on de Emociones del Usuario, 2014.

[8] E. A. Lara, L. Codigo, H. Alejandro, Q. Cruz, and E. Lara L´evano, Sistema de reconocimiento de gestos faciales captados a trav´es de c´amaras para analizar el nivel de satisfacci´on de clientes en restaurantes., 2019.

[9] G. O. and S. O., Dise˜no de un Sistema de Reconocimiento de rostros aplicando inteligencia y visi´on artificial., 2014.

[10] S. Roy., Face emotion recognition with EfficientNetB2., 2021.

[11] P. Kaur, K. Krishan, S. K. Sharma, and T. Kanchan, Facial-recognition algorithms: A literature review. [Online]. Available: https://doi.org/10.1177/0025802419893168

[12] M. Collins and S. Robert, A Generalization of Principal Component Analysis to the Exponential Family. [Online]. Available: https://www.researchgate.net/publication/2407485 A Generalization of Principal Component Analysis to the Exponential Family

[13] I. T. Jolliffe, Principal Component Analysis. [Online].
Available: https://doi.org/10.1007/b98835

[14] M. Turk and A. Pentland, Eigenfaces for Recognition. [Online].
Available: https://doi.org/10.1162/jocn. 1991.3.1.71

[15] G. L. Baume, Breve introducci´on a Google Colab., 2021.
[16] J. M. Uriarte, Google Drive., 2020.
[17] P. S. Foundation, El tutorial de Python., 2022.
[18] La librería Numpy. [Online].
Available: https://aprendeconalf.es/docencia/python/manual/numpy/
[19] GeeksforGeeks, M´odulo OS en Python con ejemplos., 2022.
[20] A. Mordvintsev and A. R. K., Tutoriales de Introducci´on a OpenCV-Python., 2022.
[21] Scikit, Tutorial de aprendizaje de Scikit. [Online].
Available: https://www.tutorialspoint.com/scikit learn/ index.ht


LaSalle
Universidad
Perú