2023 WICC
UNNOBA | JUNÍN

LIBRO DE ACTAS

XXV Workshop de Investigadores en Ciencias de la Computación

Buenos Aires, abril del 2023
UNNOBA – Red de Universidades de Carreras de Informática RedUNCI

RedUNCI ITT Instituto Transferencia Tecnología ESCUELA DE TECNOLOGÍA UNNOBA 1983 - 2023 40 años EN DEMOCRACIA Universidad Nacional Noroeste • Buenos Aires

Datos de la edición

Escuela de Tecnología

Directora: Sarobe, Mónica
Secretario Académico: Cicerchia, Lucas Benjamin

Instituto de Investigación y Transferencia en Tecnología – ITT

Director: Ramón, Hugo
Investigador: Esnaola, Leonardo
Investigador: Tessore, Juan Pablo

Red de Universidades de Carreras de Informática – RedUNCI

Representante: Russo, Claudia

Edita

CEDi Centro de Edición y Diseño. UNNOBA
Callao 289 3.° piso, CP. 1022
Tel: 54 -11-535-31520. Ciudad Autónoma
de Buenos Aires, Argentina

Diseño y diagramación de tapa e interior:
CEDi | cedi@unnoba.edu.ar

Fecha de catalogación:

Queda hecho el depósito que marca la ley 11723
Reservados todos los derechos

Abril | 2023

‣ Diseño : Reflexiones sobre el hacer en el territorio, 2018- 2020 / Maria de las Mercedes Filpe ... [et al.] ; compilación de Alejandra Mizrahi ... [et al.]. - 1a ed. - Junín : Universidad Nacional del Noroeste de la Provincia de Buenos Aires. UNNOBA, 2022.
Libro digital, PDF - (Investigación)

Archivo Digital: descarga
ISBN 978-987-3724-64-0

1. Diseño Indumentaria. 2. Diseño Textil. 3. Educación. I. Filpe, Maria de las Mercedes. II. Alejandra Mizrahi, comp.
CDD 746.92

ISBN 978-987-3724-64-0

Esta obra está bajo una licencia de Creative Commons Reconocimiento - NoComercial - SinObraDerivada 4.0 Internacional.

Índice

ARSO – Arquitectura, Redes y Sistemas Operativos	**Pág (12-57)**
Desafíos y buenas prácticas para pequeños proveedores de Internet zonales.	**Pág (13-16)**
Migración controlada de procesos en sistemas distribuidos.	**Pág (17-21)**
Estudios del aporte de tecnologías como BLOCKCHAIN, NFT y VPN a la seguridad en plataformas de IOT.	**Pág (22-26)**
IoT for Vineyard Posts.	**Pág (27-31)**
IoT-based energy monitoring system, to deploy an AMI.	**Pág (32-36)**
In a Micro Smart Grid environment.	**Pág (37-41)**
Utilización Eficiente y Sostenible de Recursos en Clúster de bajo costo.	**Pág (42-46)**
Tecnologías de Smart IoT para la solución de problemas en el medio socio productivo.	**Pág (47-52)**
Despliegue de aplicaciones contenerizadas: un caso de implementación basado en crane.	**Pág (53-57)**
ASI – Agentes y Sistemas Inteligentes	**Pág (58-156)**
Interfaces de Programación de Aplicaciones orientadas a la Integración de Sistemas Judiciales.	**Pág (59-63)**
Organizaciones inteligentes asistidas por tecnologías 4.0.	**Pág (64-68)**
Aprendizaje Automático para Nodos Inteligentes.	**Pág (69-71)**
Depuración de bases de datos de segunda sección del Boletín oficial de la república argentina mediante Aprendizaje de máquina.	**Pág (72-76)**
Una herramienta de Inteligencia Artificial para el modelado de distintos escenarios con la determinación del Índice de Calidad del Agua (ICA) para aplicar a ríos urbanos de Argentina.	**Pág (77-81)**
Movilidad inteligente en una ciudad sostenible.	**Pág (82-86)**
Asistente virtual con comportamiento definible de consultas referidas a la legislación argentina para personas con discapacidad.	**Pág (87-89)**
Automatización robótica de procesos en las pymes.	**Pág (90-94)**
Clasificación de partículas de arena a través de redes neuronales convolucionales.	**Pág (95-100)**
Resolución de problemas regionales NP-duros usando técnicas de optimización.	**Pág (101-105)**
Confección de dataset de patentes de automóviles argentinos para entrenamiento de redes neuronales.	**Pág (106-110)**
Metaheurísticas aplicadas al alineamiento múltiple de secuencias de ADN y/o de aminoácidos.	**Pág (111-114)**
Aplicación de inteligencia computacional y computación de alto desempeño en el desarrollo de un modelo de predicción de las condiciones predisponentes al quemado del arroz (pyricularia oryzae).	**Pág (115-119)**
Lenguaje natural aplicado a los sistemas conversacionales para el aprendizaje de lenguas extranjeras.	**Pág (120-124)**
Tecnologías Semánticas para el desarrollo de Agentes Inteligentes: Generación de comentarios a partir de código fuente.	**Pág (125-128)**
Deep Learning para Visión por Computadora.	**Pág (129-133)**
Inteligencia artificial como herramienta para innovar y dinamizar procesos.	**Pág (134-139)**
Detección de Ransomware en Blockchains.	**Pág (140-142)**
Desarrollo de soluciones tecnológicas de bajo costo a problemáticas en el sector hortícola.	**Pág (143-147)**
Desarrollo de herramientas para generación de material de entrenamiento y análisis de textos estructurados.	**Pág (148-152)**
Sistema de predicción de incendios forestales.	**Pág (153-156)**

BDDM – Bases de Datos y Minería de Datos	**Pág (157-216)**
Propuesta de técnicas de validación para la calidad de datos abiertos e identificación de patrones para predicciones con Machine Learning: Segunda Parte.	**Pág (158-162)**
Knowledge Graphs para accesibilidad y reutilización de activos digitales del Mar Argentino.	**Pág (163-166)**
Descubrimiento de patrones de comportamiento vinculados al abandono en la Universidad Nacional de Luján mediante la aplicación de técnicas de aprendizaje automático.	**Pág (167-173)**
Desarrollo de soluciones de inteligencia y analítica de negocios.	**Pág (174-177)**
Rediseño de la Plataforma Web MultiDB, Cuyo Propósito es el Aprendizaje de Lenguaje SQL, con Acceso a Distintos Motores de Base de Datos.	**Pág (178-182)**
Sistemas inteligentes en el uso de aplicaciones de bioinformática y sistemas embebidos.	**Pág (183-187)**
Datos no Estructurados: Indexación, Búsquedas y Aplicaciones	**Pág (188-192)**
Administración y Recuperación de Información en Bases de Datos Masivas.	**Pág (193-197)**
Diversas Aplicaciones de Bases de Datos Espaciales y Espacio Temporales.	**Pág (198-202)**
Bases de Datos no Convencionales.	**Pág (203-207)**
Comparación de métodos y métricas para la predicción de series temporales aplicados a las ventas masivas de productos minoristas.	**Pág (208-212)**
Índices de dominio para identificación eficiente mediante código de iris.	**Pág (213-216)**
CGIV – Computación Gráfica, Imágenes y Visualización	**Pág (217-261)**
Testing de visualizaciones.	**Pág (218-221)**
PACS Cloud Open Source.	**Pág (222-226)**
Análisis visual de movimientos oculares en entornos dinámicos.	**Pág (227-231)**
Visualización comparativa: estudio y diseño de visualizaciones orientadas a la comparación de conjunto de datos.	**Pág (232-236)**
Tecnologías inmersivas y visualización situada aplicadas a geociencias.	**Pág (237-241)**
Visualización de datos multidimensionales procedentes de las geociencias.	**Pág (242-246)**
Análisis visual guiado para el estudio de datos geoquímicos.	**Pág (247-251)**
Industrias Inteligentes: Detección de uso de Elementos de Protección Personas en una Planta de Reciclaje.	**Pág (252-256)**
La importancia de la realidad extendida en el metaverso.	**Pág (257-261)**
IEI – Innovación en Educación en Informática	**Pág (262-331)**
Adopción de verilog en materias de diseño digital.	**Pág (263-267)**
Aplicación con realidad aumentada para el aprendizaje de los conceptos de programación.	**Pág (268-272)**
Innovación Tecnológica y Metodológica en Educación.	**Pág (273-277)**
Evaluación de la Calidad de Objetos de Aprendizaje, para la enseñanza de la Informática, desde el Enfoque de los estudiantes. Medición del significado afectivo.	**Pág (278-282)**
Marco para desarrollar nuevas unidades de proyectos e incorporar docentes y alumnos en actividades de investigación.	**Pág (283-287)**
Herramientas colaborativas multiplataforma para el aprendizaje basado en competencias en la ingeniería de software.	**Pág (288-295)**
El portafolio como metodología de trabajo en un grupo de investigación de infraestructura it.	**Pág (296-300)**
Enseñar programación en la universidad en tiempos de gran demanda de programadores.	**Pág (301-305)**
Eye tracking para la enseñanza de la programación.	**Pág (306-310)**
Demanda laboral y competencias en inteligencia artificial dentro de carreras grado en ciencias aplicadas.	**Pág (311-315)**

Proyecto Esport UNLP Diplomatura en Deportes Electrónicos.	Pág (316-321)
Programación y Autismo: una arquitectura epistémica centrada en Ciencia Ciudadana.	Pág (322-326)
Desarrollo de Recursos Educativos para la Enseñanza de las Ciencias de la Computación en la Escuela Secundaria.	Pág (327-331)
IS – Ingeniería en Software	**Pág (332-481)**
Principios esenciales de la ingeniería de software en ámbito público.	Pág (333-337)
Catálogo y recomendaciones de recursos (métodos) para facilitar experiencias de Diseño Participativo usando Design Thinking.	Pág (338-342)
Análisis de viabilidad del despliegue de sistemas de software en PyMES.	Pág (343-347)
Métodos, técnicas y herramientas para mejorar y evaluar la calidad de requisitos en proyectos ágiles de software.	Pág (348-352)
Métodos, técnicas y herramientas para mejorar y evaluar la calidad de la capacidad de producción de entregables dentro de un proceso ágil de desarrollo de software	Pág (353-357)
Creación de un observatorio tecnológico de la gestión del conocimiento en la ciudad de buenos aires.	Pág (358-362)
Aprovisionamiento automático en la nube mediante la Ingeniería Dirigida por Modelos.	Pág (363-367)
Hacia la evaluación de la calidad de datos abiertos.	Pág (368-372)
Aplicaciones Móviles, Realidad Virtual y Realidad Aumentada.	Pág (373-376)
Evaluación de las políticas de protección de datos personales en las iniciativas de apertura de datos en el ámbito de la Universidad Nacional de Catamarca.	Pág (377-381)
Desarrollo de Computadoras de a Bordo para Misiones Satelitales de Órbita Baja.	Pág (382-386)
Enfoques y Tendencias en el Desarrollo de Aplicaciones para Dispositivos Móviles.	Pág (387-392)
Construcción de colecciones de proyectos de software para estudios empíricos.	Pág (393-397)
Buenas prácticas de calidad de software en el ámbito público.	Pág (398-402)
Evaluación de lenguajes de modelado gráficos de procesos de negocio colaborativos utilizando un método multicriterio.	Pág (403-407)
Diseño de un marco de trabajo para la gestión del conocimiento de productos de software en evolución.	Pág (408-410)
Aspectos de ingeniería de software, bases de datos relacionales, y bases de datos NOSQL para el desarrollo de sistemas de software híbridos.	Pág (411-417)
Desarrollo de Software Público: Un aporte para la mejora de la calidad.	Pág (418-422)
Cultura, turismo y tecnología. Experiencias turísticas significativas para circuitos históricos de caleta Olivia.	Pág (423-427)
Desarrollo y evaluación de productos web y móvil centrados en la experiencia de usuario.	Pág (428-433)
Evaluación de la Sostenibilidad en un Modelo de Calidad del Software.	Pág (434-438)
CoVaMaT: Modelo conceptual de una herramienta para el soporte a la gestión de variedad en Sistemas Big Data.	Pág (439-443)
Proceso de validación de requerimientos aplicando técnicas de procesamiento de lenguaje natural en un entorno colaborativo.	Pág (444-448)
Computación afectiva aplicada a la interacción con Interfaces Gráficas de Usuario.	Pág (449-453)
Calidad de información en comunidades virtuales.	Pág (454-458)
Accesibilidad web. Estrategias y aplicaciones.	Pág (459-463)
Estimación del tamaño de un sistema mediante el análisis de diagramas UML y la aplicación de métodos basados en puntos de función.	Pág (464-467)
Gobernanza de Sistemas y Tecnologías de la Información (SI/TI) en el ámbito Universitario.	Pág (468-474)
Transformación Digital – Propuestas de solución para las Ciudades Inteligentes Sostenibles, la Gobernanza Digital y las Buenas prácticas y calidad.	Pág (475-480)

ISS – Innovación en Sistemas de Software	**Pág (481-573)**
Estado actual de la industria 4.0 en las PyMES del partido de Hurlingham.	**Pág (482-486)**
Sistema de recomendación ubicuo para contribuir a la disminución del estrés académico en estudiantes universitarios.	**Pág (487-490)**
El avance en la madurez digital de las pymes a través de la transformación digital.	**Pág (491-495)**
Tecnologías de la información facilitadoras para la interoperabilidad de software en Gobierno Abierto: Segunda Parte.	**Pág (496-500)**
Desarrollo de una app móvil de Nueva Generación.	**Pág (501-504)**
Soluciones IoT desde Smart Cities a Blockchain.	**Pág (505-510)**
Diseñar una solución de identidad auto-gestionada para acceso a servicios de calidad con redes Blockchain multipropósito en la Universidad Nacional de Río Negro.	**Pág (511-515)**
Aplicación de geofencing para el cuidado de adultos mayores mediante un hardware dedicado.	**Pág (516-519)**
Programa integrador de proyectos de Software con transferencia a la comunidad.	**Pág (520-523)**
Innovación en las Ciudades mediante las TIC Análisis de implementaciones en Ciudades Inteligentes.	**Pág (524-528)**
Aplicación de técnicas de inteligencia artificial en la prevención de problemas de salud mental.	**Pág (529-533)**
Soluciones de IOT: ámbitos de aplicación y desafíos.	**Pág (534-538)**
Machine learning y deep learning en la interpretación del lenguaje de señas.	**Pág (539-543)**
Diseño de una estrategia para la gestión del tiempo en la comunicación de corrección de fallas en el mantenimiento industrial.	**Pág (544-548)**
Implementación de una BCI mediante el diseño y desarrollo de Serious Games para la rehabilitación de pacientes neurológicos.	**Pág (549-553)**
La Inteligencia Artificial y su aplicación a los servicios de Internación Domiciliaria.	**Pág (554-558)**
Abordajes científico tecnológicos basados en el uso de vants.	**Pág (559-563)**
Innovación tecnológica a través de la hiper automatización.	**Pág (564-568)**
Tecnologías de Datos Espaciales, Visualización y Realidad Virtual.	**Pág (569-573)**
PDP – Procesamiento Distribuido y en Paralelo	**Pág (574-636)**
Paralelizacion de algoritmos y evaluacion de rendimiento en Plataformas de cómputo de altas prestaciones.	**Pág (575-580)**
Análisis y modelado de sistemas utilizando Cloud Computing, IaaS privados y públicos.	**Pág (581-585)**
Estrategias de modelado y simulación en sistemas de HPC y salud.	**Pág (586-590)**
Arquitecturas Multiprocesador en HPC: Software de Base, Modelos y Aplicaciones.	**Pág (591-596)**
Arquitecturas Edge-Fog-Cloud en Procesamiento Distribuido. Aspectos de Eficiencia y Resiliencia.	**Pág (597-602)**
Soporte Serverless para Aplicaciones Móviles de Nueva Generación.	**Pág (603-607)**
Predicción de fenómenos naturales mediante paralelismo, algoritmos evolutivos y búsqueda por novedad.	**Pág (608-612)**
Simulación y análisis eficiente de grandes volúmenes de datos.	**Pág (613-617)**
Autoescalado basado en Aprendizaje Profundo por Refuerzo de Workflows Científicos en la Nube.	**Pág (618-622)**
Análisis de rendimiento de aplicaciones y sistemas complejos mediante computación paralela y técnicas de optimización de recursos.	**Pág (623-627)**
Técnicas de Recuperación y Procesamiento de Grandes Volúmenes de Datos Sísmicos. Un Repositorio Público Basado en Serverless.	**Pág (628-631)**
Aplicaciones para la salud y métricas de rendimiento en sistemas paralelos y distribuidos.	**Pág (632-636)**

PSSTR – Procesamiento de Señales y Sistemas en Tiempo Real Pág (637-655)

Avances en el desarrollo de un prototipo experimental de segmento terreno Satelital multiplataforma-multimisión UGS UNLaM. Pág (638-641)

Partición de Visual SLAM para su uso con celulares. Pág (642-646)

Implementación y Simulación de Sistemas de Tiempo Real. Pág (647-650)

Captura de variables para alerta temprana de incendios forestales y su almacenamiento para integración a sistemas de predicción mediante utilización de redes inalámbricas de sensores. Pág (651-655)

RCCI – Redes de Cooperación Científica Internacionales Pág (656-666)

Maestría en Gestión y Tecnología de Ciudades Inteligentes Resultado del Proyecto CAP4CITY. Pág (657-660)

Avances del Consorcio para la colaboración en I+D+I en Temas de Cloud Computing, Big Data y Emerging Topics (CCC-BD&ET). Pág (661-666)

SI – Seguridad Informática Pág (667-726)

Estudio de las propiedades criptológicas de los Tokens No Fungibles Inteligentes para asegurar dispositivos de Internet de las Cosas. Pág (668-673)

Análisis de las Herramientas para Realizar Pruebas Estáticas de Seguridad de las Aplicaciones. Pág (674-679)

Tecnología Blockchain Aplicada a la Ciberdefensa. Pág (680-683)

Aprendizaje por Refuerzo Aplicado al Etiquetado de Tráfico de Red. Pág (684-688)

Criptografía Liviana y Ciberseguridad aplicados a Sistemas Ciberfísicos. Pág (689-693)

Software para recolección de evidencias digitales rápidas en sistemas Windows. Pág (694-698)

Protección de activos de software. Pág (699-703)

Técnicas para Incrementar la Seguridad en Web Services basados en WSDL. Pág (704-708)

Avances sobre la Especificación Integral del Sistema OTP-Vote Orientada a su Implementación. Pág (709-713)

Criptología Maliciosa para la Ciberdefensa. Pág (714-718)

Métodos, Técnicas y Herramientas para la Protección de Sistemas de Software. Pág (719-723)

La bockchain y la certificación de documentos en Argentina. Pág (724-726)

TD – Tesis Doctorales Pág (727-775)

Generalización del modelado de cadencias de tecleo con contextos finitos para su utilización en ataques de presentación y canal lateral. Pág (728-737)

Cálculo Científico Distribuido sobre Clientes Móviles Indeterminados en redes MANETs. Pág (738-747)

DepProMod: Modelo de Proceso de Despliegue de Sistemas de Software. Pág (748-757)

Tesis Doctoral. Métricas de calidad para validar los conjuntos de datos abiertos públicos gubernamentales. Pág (758-767)

Modelo de madurez para servicios de gobierno electrónico en el ámbito universitario. Pág (768-775)

TIAE – Tecnología Informática Aplicada en Educación Pág (776-897)

Tecno pedagogía en el Proceso Formativo. Pág (777-781)

Estrategias de accesibilidad para una educación inclusiva en la UNSAdA. Pág (782-786)

Diseño de herramientas tecnológicas en Modelos educativos híbridos. Pág (787-791)

Recursos educativos accesibles. Estudio de distintas técnicas y herramientas basadas en inteligencia artificial. Pág (792-797)

Experiencia de Usuario en Sistemas Interactivos usando reconocimiento de Emociones. Pág (798-802)

Un bot virtual en Moodle para asistir y complementar actividades académicas en espacios de educación superior.	**Pág (803-807)**
Interacción Persona Ordenador y Educación. Desarrollos y experiencias.	**Pág (808-813)**
Realidad aumentada en contextos educativos y su relación con el rendimiento académico universitario.	**Pág (814-815)**
Tecnologías digitales aplicadas al escenario educativo. Desarrollos y experiencias.	**Pág (816-820)**
Tecnologías disruptivas y aulas emergentes en el nivel medio y superior en el contexto de la UNRN.	**Pág (821-825)**
Desarrollo y aplicación de laboratorios virtuales destinados a educación de grado y pregrado.	**Pág (826-830)**
La mejora del aprendizaje en AVAs aplicando analítica del aprendizaje.	**Pág (831-835)**
La incorporación de la Ciencia de Datos como estrategia transversal para refozar la enseñanza de la programación en el aula.	**Pág (836-840)**
Realidad Virtual y Educación.	**Pág (841-845)**
Diagnóstico de competencias genéricas universitarias, a través de software business intelligence.	**Pág (846-850)**
Revisión de alternativas moodle para la extracción de indicadores de autorregulación del aprendizaje.	**Pág (851-855)**
Realidad virtual para la enseñanza de portugués: Una guía para la implementación de contenido educativo en estudiantes de nivel medio.	**Pág (856-860)**
Definición de Lenguajes Específicos del Dominio para la Generación de Sitios Web de Instituciones Escolares.	**Pág (861-865)**
Definición de Lenguajes Específicos del Dominio para la Generación de Sitios Web de Instituciones Escolares.	**Pág (866-869)**
Minisatélite CANSAT para la enseñanza de las ciencias experimentales en diferentes contextos educativos.	**Pág (870-874)**
Calidad educativa: instrumentos para alcanzar la mejora continua de la educación superior.	**Pág (875-879)**
Evaluación del desempeño docente en el nivel superior.	**Pág (880-884)**
Software para gamificar terapias de rehabilitación AD-HOC.	**Pág (885-890)**
Tecnologías innovadoras aplicadas en contextos educativos	**Pág (891-894)**
Alfadatizando: innovación en el proceso de enseñanza aprendizaje en escuelas secundarias usando visualización de datos para desarrollar pensamiento computacional en materias de Humanidades y Ciencias Sociales.	**Pág (895-897)**
Autores:	**Pág (898-901)**

DEPURACIÓN DE BASES DE DATOS DE SEGUNDA SECCIÓN DEL BOLETÍN OFICIAL DE LA REPÚBLICA ARGENTINA MEDIANTE APRENDIZAJE DE MAQUINA

Nestor A. Balich, Franco A. Balich, Hugo Fraga

CAETI - Centro de Altos Estudios en Tecnología Informática
Universidad Abierta Interamericana. Informática (UAI)
Montes de Oca 745. Ciudad Autónoma de Buenos Aires, Argentina.

{nestor.balich, francoadrian.balich}@uai.edu.ar

hfraga@boletinoficial.gob.ar

RESUMEN

El Boletín Oficial de la República Argentina es una fuente importante de información para empresas, instituciones y particulares, pero la gran cantidad de información publicada diariamente hace que la depuración de bases de datos sea un proceso complejo y costoso. Se hipotetiza que el uso de un modelo de inteligencia artificial basado en aprendizaje de maquina (ML) puede aprender a clasificar y catalogar los avisos comerciales de manera eficiente y autónoma. La aplicación de técnicas de ML en la clasificación de textos ha evolucionado significativamente, especialmente en el uso de modelos de aprendizaje profundo en áreas como la identificación de noticias falsas y la detección de spam. También se han realizado estudios sobre el uso de técnicas de ML en la identificación y corrección de errores en bases de datos, incluyendo la corrección de errores en imágenes médicas y la identificación de avisos comerciales ilegales en la web. Se propone el desarrollo de un modelo de aprendizaje de máquina para la clasificación y catalogación de los avisos comerciales de la segunda sección del Boletín Oficial de la República Argentina, lo que podría representar un avance significativo en la depuración de bases de datos en este ámbito.

Palabras clave: Depuración, Base de datos, Aprendizaje de Maquina, Inteligencia Artificial, Aprendizaje Supervisado, Clasificación

CONTEXTO

Esta línea de I+D forma parte de los proyectos radicados en el Centro de Altos Estudios en Tecnología informática (CAETI) de la Universidad Abierta Interamericana (UAI). En este proyecto participan docentes, alumnos e investigadores enmarcado en los proyectos de transferencia tecnología del laboratorio de robótica física (LRF). Las líneas de investigación sobre inteligencia dentro de los proyectos con financiamiento y duración a 2 años.

1. INTRODUCCIÓN

La segunda sección del Boletín Oficial de la República Argentina es un importante medio de difusión de información para empresas, instituciones

y particulares. Sin embargo, la cantidad de información que se publica diariamente en esta sección hace que la depuración de bases de datos sea un proceso complejo y costoso en términos de tiempo y recursos. En este contexto, surge la hipótesis de que es posible desarrollar un modelo de inteligencia artificial basado en ML capaz de aprender a clasificar los avisos comerciales y luego catalogarlos para obtener los avisos de manera eficiente y de forma totalmente autónoma.

La aplicación de estas técnicas en la clasificación de textos es un área de investigación en constante evolución, y su aplicación en la depuración de bases de datos no es una excepción. En particular, la utilización de modelos de aprendizaje profundo ha permitido mejorar significativamente la capacidad de clasificación de textos en diversas áreas, como la identificación de noticias falsas o la detección de spam en correos electrónicos.

En el ámbito de la depuración de bases de datos, también se han realizado estudios sobre la utilización de técnicas de ML para la identificación y corrección de errores *"el uso de Machine Learning para mejorar la eficiencia y la precisión de la limpieza de datos y la consideración de los efectos de la limpieza de datos en análisis estadístico"* Chu [1]

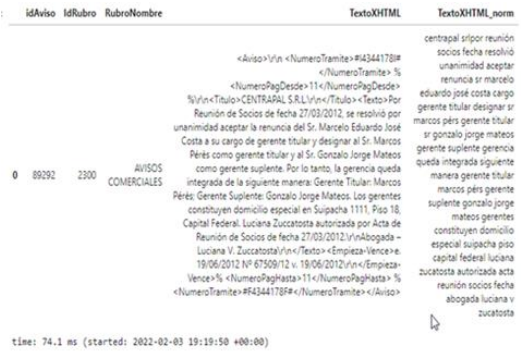
Los problemas de calidad de los datos son complejos de resolver y existente gran cantidad de técnicas y herramientas para lograrlo Gudivada [2] y que debimos evaluar al seleccionar los modelos a utilizar en el presente trabajo.

Otro estudio interesante donde se desarrolló un modelo para la detección de publicidad ilegal en vallas publicitarias basado en el aprendizaje automático (FIBAD) Huxiao[3]. Permitió demostrar que los resultados obtenidos por el modelo eran capaces de identificar avisos ilegales con una alta precisión y eficiencia.

En este trabajo, se propone el desarrollo y validación de un modelo de inteligencia artificial basado en ML con aprendizaje supervisado para la clasificación y catalogación de los avisos comerciales de la segunda sección del Boletín Oficial de la República Argentina. Se espera que este modelo permita obtener los avisos de manera eficiente y autónoma, lo cual representaría un avance significativo en la depuración de bases de datos en este ámbito. *"Los datos son uno de los desafíos perennes en el análisis de datos, y no hacerlo puede resultar en análisis inexactos y decisiones poco confiables*" Khayyat [4].

Luego de realizar la (1) adquisición y depuración de datos se realizó el (2) proceso de análisis exploratorio de datos (EDA), donde los datos se examinan con estadísticas descriptivas simples para determinar si existen problemas de datos, como podemos observar detalladamente en el libro *"Data Science in the Database: Using SQL for Data Preparation"* Badia [5]. Continuando con la identificación de los datos representativos obtenidos de cinco tablas (id del tipo de aviso, texto del aviso, texto del tema, id de registro, fecha de ingreso, fecha de publicación, número de boletín) generando un archivo avisos2da.csv con un total de 770.000 registros. En el siguiente paso consistió en la (3) normalización de los datos eliminando los registros con

republicaciones figura 1, pues el mismo aviso debe ser republicado varios días y con mismo texto. Para fines del entrenamiento solo necesitamos una sola muestra, reduciéndose el dataset a 550.000 registros.



	idAviso	IdRubro	RubroNombre	TextoXHTML	TextoXHTML_norm
0	89292	2300	AVISOS COMERCIALES	<Aviso>\r\n <NumeroTramite>#F4344178F# </NumeroTramite> % <NumeroPagDesde>11</NumeroPagDesde> %\r\n<Titulo>CENTRAPAL S.R.L\r\n</Titulo><Texto>Por Reunión de Socios de fecha 27/03/2012, se resolvió por unanimidad aceptar la renuncia del Sr. Marcelo Eduardo José Costa a su cargo de gerente titular y designar al Sr. Marcos Pérès como gerente titular y al Sr. Gonzalo Jorge Mateos como gerente suplente. Por lo tanto, la gerencia queda integrada de la siguiente manera: Gerente Titular: Marcos Pérès; Gerente Suplente: Gonzalo Jorge Mateos. Los gerentes constituyen domicilio especial en Suipacha 1111, Piso 18, Capital Federal. Luciana Zuccatosta autorizada por Acta de Reunión de Socios de fecha 27/03/2012.\r\nAbogada – Luciana V. Zuccatosta\r\n</Texto><Empieza-Vence>e. 19/06/2012 Nº 67509/12 v. 19/06/2012\r\n</Empieza-Vence>% <NumeroPagHasta>11</NumeroPagHasta> % <NumeroTramite>#F4344178F#</NumeroTramite></Aviso>	centrapal srlpor reunión socios fecha resolvió unanimidad aceptar renuncia sr marcelo eduardo josé costa cargo gerente titular designar sr marcos pérs gerente titular sr gonzalo jorge mateos gerente suplente gerencia queda integrada siguiente manera gerente titular marcos pérs gerente suplente gonzalo jorge mateos gerentes constituyen domicilio especial suipacha piso capital federal luciana zucatosta autorizada acta reunión socios fecha abogada luciana v zucatosta

time: 74.1 ms (started: 2022-02-03 19:19:50 +00:00)

Figura 1: Registro base

Se procedió a la (4) agrupación por categorías de tipo de aviso figura 2, detectando algunos tipos de avisos con menos de 20 ocurrencias que fueron descartados por no ser suficientes para el entrenamiento. Fijando la cantidad de avisos representativos mínima en 450 para poder extraer una muestra aleatoria y balanceada de cada categoría. Pues *"Los conjuntos de datos desequilibrados afectan negativamente el rendimiento de los algoritmos de aprendizaje automático"* Mooijman[6].

RubroNombre	
AVISOS COMERCIALES	180641
BALANCES	527
CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEBRAS. OTROS	30172
CONSTITUCION SA	35408
CONSTITUCION SAS	4710
CONTRATO SRL	47270
CONVOCATORIAS	23162
ESTATUTO OTRAS SOCIEDADES	146
ESTATUTO SCA	6
INFORMACION Y CULTURA	481
MODIFICACIONES SRL	26441
PARTIDOS POLITICOS	4538
REFORMA OTRAS SOCIEDADES	452
REFORMA SA	35664
REFORMA SCA	641
REMATES COMERCIALES	1248
REMATES JUDICIALES	8183
SUCESIONES	103539
TRANSF. FONDO DE COMERCIO	1994

Figura 2. Agrupación por categoría

Se procedió a establecer una (5) base de referencia con modelo de aprendizaje automáticos, procesando el set de entrenamiento normalizado (7.000 avisos) con AutoNLP, que lo agrupo y clasificó por tipos societarios, dividiendo el set en dos grupos (Entrenamiento, 70% y Test, 30%).

(6) Creamos 7 modelos (Naïve Bayes NB, GS, Logistic regression LogR, linear regression LinR, random forest classifier RFC, K-Nearest Neighbor KNC, Support Vector Machines SVC) candidatos chowdhury[7] en su respectiva notebook jupyter. Cada uno de ellos fue entrenado con el mismo 70% del set de 7.000 avisos y luego validado contra el 30% obtenido los resultados expresados en la figura 3.

	Model	Accuracy	f1	precision	recall
0	NB	86.90	86.56	88.04	86.90
1	GS	93.77	93.70	93.88	93.77
2	LogR	91.73	91.67	91.93	91.73
3	LinR	93.68	93.62	93.74	93.68
4	RFC	91.45	91.33	91.65	91.45
5	KNC	84.76	84.61	85.42	84.76
6	SVC	93.68	93.62	93.74	93.68

Figura 3. Resultados de modelos de IA

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

Los ejes principales del trabajo son:

- Implementar un modelo de IA que aprenda a identificar avisos de 3ra sección.
- Implementar un modelo que valide y permita depurar la base de datos existente.
- Evidencia la performance y vialidad del empleo de IA en el

proceso de depuración de base de datos clasificadas.

3. RESULTADOS OBTENIDOS/ESPERADOS

Una vez probado en distintos modelos, se define el de mejor eficacia y se procede a aplicar a los 550.000 avisos disponibles en la 2da. Sección. Obteniendo una eficacia de entre el 84,76% y 93,77%.
Se obtuvo un total de 30.000 avisos sobre los 550.000 catalogados para revisión, de los cuales encontramos que sobre una muestra de 300 avisos 200 estaban mal clasificados y 100 no tenían el formato correspondiente o faltaban datos sobre el contenido del aviso.

Concluimos que ante las sucesivas migraciones de la base de datos a lo largo 20 años, se ha cambiado el formato y estructura de los avisos, pese a esto al ser un formato legal los principales indicadores están presentes permitiendo al modelo de aprendizaje supervisado aprender y clasificar con alto grado de asertividad (superiores al 84%).

El proceso total de trabajo desde la adquisición de los datos, depuración y entrenamiento demanda 3 días de trabajo, contra un proceso manual que se estima en más de un año si realizara de forma manual por personal de publicaciones con una dotación de 10 empleados.

El resultado es alentador y ya con el set de datos a verificar el personal de publicaciones estima un mes de trabajo para verificar y corregir las inconsistencias.

Como proyecto a futuro se contempla incorporar los modelos de IA a las aplicaciones existentes, tanto para validad el ingreso de los avisos, como para contar con una herramienta de validación histórica en tiempo real. Y aplicar los modelos de aprendizaje a 1ra sección de la base de datos Boletín Oficial que supera los 2.000.000 de avisos.

También una nueva línea de investigación que se desprende del presente trabajo, sobre el análisis mediante inteligencia artificial de la base de datos legales. Como lo expresan varios autores "*la clasificación automatizada de textos legales es un tema de investigación destacado en el campo legal. Sienta las bases para construir un sistema legal inteligente."* Haihua [8]. Enmarcados dentro de los proyecto de investigación y avances de los sistemas con inteligencia artificial en el sistema legal Argentino y su aplicabilidad Dobratinich[9] y Dobratinich [10].

4. FORMACION DE RECURSOS HUMANOS

El equipo está formado por 1 docente Magister en tecnología informática graduado de la Universidad Abierta Interamericana (UAI), 1 alumnos de 5to año de ingeniería en sistemas (UAI) y un investigador del laboratorio de robótica física (LRF-UAI).

Este proyecto también cuenta con la participación de colaboradores externos del área de sistemas y publicaciones de boletín oficial.

5. BIBLIOGRAFIA

[1]Chu, Xu & Ilyas, Ihab & Krishnan, Sanjay & Wang, Jiannan. (2016). Data Cleaning: Overview and Emerging

Challenges. 2201-2206. 10.1145/2882903.2912574. https://dl.acm.org/doi/proceedings/10.1145/2882903

[2] Gudivada, Venkat & Apon, Amy & Ding, Junhua. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software. 10. 1-20. https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations

[3] Huxiao Liu, Lianhai Wang, Weinan Zhang, Wei Wang (2019). An Illegal Billboard Advertisement Detection Framework Based on Machine Learning https://doi.org/10.1145/3358528.3358549

[4] Khayyat, Zuhair & Ilyas, Ihab & Jindal, Alekh & Madden, Samuel & Ouzzani, Mourad & Papotti, Paolo & Quiané-Ruiz, Jorge-Arnulfo & Tang, Nan & Yin, Si. (2015). BigDansing: A System for Big Data Cleansing. 10.1145/2723372.2747646. https://www.researchgate.net/publication/274256636_BigDansing_A_System_for_Big_Data_Cleansing

[5] Badia Antonio (2023). Data Science in the Database: Using SQL for Data Preparation . University of Louisville, USA - DOI: 10.4018/978-1-7998-9220-5.ch069

[6] Paul Mooijman, Cagatay Catal, Bedir Tekinerdogan, Arjen Lommen, Marco Blokland. The effects of data balancing approaches: A case study,Applied Soft Computing, https://doi.org/10.1016/j.asoc.2022.109853.

[7] Shovan Chowdhury, Marco P. Schoen (2020), Research Paper Classification using Supervised Machine Learning Techniques - Intermountain Engineering, Technology and Computing (IETC) – IEEE https://ieeexplore.ieee.org/document/9249211

[8] Haihua Chen, Lei Wu, Jiangping Chen, Wei Lu, Junhua Ding (2022). A comparative study of automated legal text classification using random forests and deep learning, Information Processing & Management,Volume 59, Issue 2, 2022, 102798, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2021.102798.

[9] Dobratinich Gonzalo (2022). INTELIGENCIA ARTIFICIAL Y JUSTICIA: APLICABILIDAD DE LA TECNOLOGÍA EN LAS DECISIONES JUDICIALES EN ARGENTINA. Revista Direitos Culturais. 17. 203-216. 10.20912/rdc.v17i42.761.

https://www.researchgate.net/publication/363593342_INTELIGENCIA_ARTIFICIAL_Y_JUSTICIA_APLICABILIDAD_DE_LA_TECNOLOGIA_EN_LAS_DECISIONES_JUDICIALES_EN_ARGENTINA

[10] Gonzalo Ana Dobratinich (2021)- Evaluación De La Preparación Del Sistema Judicial Para La Adopción
De Inteligencia Artificial – Universidad de San Andrés - https://repositorio.udesa.edu.ar/jspui/handle/10908/18634