ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
PROGRAMA EN MAESTRÍA EN CIENCIAS CON MENCIÓN EN TECNOLOGIA DE LA INFORMACIÓN

Tesis

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín en 2022

Para optar el grado académico de Maestro en Ciencias con Mención en Tecnología de la Información

Autor:

Even Ronald Pérez Díaz
https://orcid.org/0000-0002-9281-7418

Asesor:

Dr. Alberto Alva Arévalo
https://orcid.org/0000-0002-8392-3542

Tarapoto, Perú

2024

ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
PROGRAMA EN MAESTRÍA EN CIENCIAS CON MENCIÓN EN TECNOLOGIA DE LA INFORMACIÓN

Tesis

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022

Para optar el grado académico de Maestro en Ciencias con Mención en Tecnología de la Información

Autor:

Even Ronald Díaz
https://orcid.org/0000-0002-8392-3542

Asesor:

Dr. Alberto Alva Arévalo
https://orcid.org/0000-0002-8392-3542

Tarapoto, Perú

2024

ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
PROGRAMA DE MAESTRÍA EN CIENCIAS CON MENCIÓN EN TECNOLOGÍA DE LA INFORMACIÓN

Tesis

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022

Para optar el grado académico de Maestro en Ciencias con Mención en Tecnología de la Información

Autor:

Even Ronald Pérez Díaz

Sustentado y aprobado el 22 de mayo del 2024, por los siguientes jurados:

Presidente de Jurado
Ing. Dr. Jorge Damián Valverde Iparraguirre

Secretario de Jurado
Ing. Dr. Juan Orlando Riascos Armas

Miembro de Jurado
Ing. Dr. Elmer Ruíz Trigozo

Asesor
Ing. Dr. Alberto Alva Arévalo

Tarapoto, Perú

2024

ACTA DE SUSTENTACIÓN DE TESIS

Los Miembros del Jurado que suscriben, reunidos para estudiar y escuchar la sustentación y defensa del Trabajo de Tesis, modo presencial, presentado por:

Bach. Even Ronald Pérez Díaz

Con el asesoramiento del Ing. Dr. Alberto Alva Arévalo.

"Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martin En 2022"

Teniendo en consideración los méritos del referido trabajo, así como los conocimientos demostrados por el sustentante, lo declaramos: APROBADO

Con el calificativo (*) MUY BUENO
DIECISIETE (17)

En consecuencia, queda en condición de ser considerado APTO por el Consejo Universitario y recibir el Grado Académico de Maestro, de conformidad con lo estipulado en el Artículo 30° del Reglamento de Tesis de la Escuela de Posgrado de la UNSM.

Tarapoto, 22 de mayo de 2024.

Ing. Dr. Jorge Damián Valverde Iparraguirre
Presidente

Ing. Dr. Juan Orlando Riascos Armas
Secretario

Ing. Dr. Elmer Ruiz Trigozo
Miembro

Ing. Dr. Alberto Alva Arévalo
Asesor

(*) De acuerdo con el Artículo 40° del Reglamento General de Ciencia, Tecnología e Innovación (RG - CTI) la Universidad Nacional de San Martín - Tarapoto, estas deberán ser calificadas con términos de: BUENO, MUY BUENO, EXCELENTE, también considerar la nota

ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
PROGRAMA DE MAESTRIA EN CIENCIAS CON MENCIÓN EN TECNOLOGÍA DE LA INFORMACIÓN

Tesis

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022

Para optar el grado académico de Maestro en Ciencias con mención en Tecnología de la Información

El suscrito declara que el presente trabajo de tesis es original, en su contenido y forma.

Ejecutor
Even Ronald Pérez Díaz

Asesor
Ing. Dr. Alberto Alva Arévalo

Tarapoto, Perú

2024

Declaratoria de autenticidad

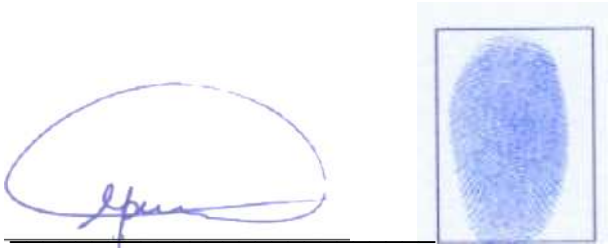
Yo, Even Ronald Pérez Díaz, identificado con DNI N° 43386240, egresado de la Escuela de Posgrado de la Universidad Nacional de San Martín, Unidad de Posgrado de la Facultad de Ingeniería de Sistemas e Informática, Programa de Maestría en Ciencias con Mención en Tecnología de la Información, con la tesis titulada: “Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022”

Declaro bajo juramento que:

1. Declaro que he redactado completamente esta tesis.
2. Todas las fuentes consultadas están debidamente citadas y referenciadas según estándares internacionales, asegurando que no he plagiado parte alguna de la tesis.
3. Esta tesis no ha sido publicada previamente ni ha sido utilizada para obtener otro grado académico.
4. Los resultados presentados son genuinos y no han sido manipulados, duplicados ni tomados de otras fuentes, representando contribuciones originales a la investigación realizada.

En caso de que considere que el trabajo contiene una falla grave, como datos fraudulentos, evidencias manipuladas, o plagio (ya sea al no citar adecuadamente las fuentes o al presentar información de otros trabajos como propia), así como falsificación (al atribuirse información e ideas de otras personas de manera incorrecta), acepto las consecuencias y sanciones que resulten de mi acción, y me someto a la normativa vigente de la Universidad Nacional de San Martín.

Tarapoto, 22 de mayo de 2024.



Even Ronald Pérez Díaz
DNI N° 43386240

Ficha de identificación

Título del proyecto Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022	**Área de investigación:** Ingeniería y Control **Línea de investigación:** Ingeniería de Sistemas y Comunicaciones **Grupo de investigación:** Automatización de Procesos y Robótica (GIAR) **Tipo de investigación:** Básica ☐, Aplicada ☒, Desarrollo experimental ☐

Autor: Even Ronald Pérez Díaz	Escuela de Posgrado Unidad de Posgrado de la Facultad de Ingeniería de Sistemas e Informática Programa de Maestría en Ciencias con mención en Tecnología de la Información https://orcid.org/0000-0002-9281-7418

Asesor: Alberto Alva Arévalo	**Dependencia local de soporte:** Facultad de Ingeniería de Sistemas e Informática Escuela Profesional de Ingeniería de Sistemas e Informática Unidad o Laboratorio Ingeniería de Sistemas e Informática https://orcid.org/0000-0002-8392-3542

Dedicatoria

A mis amados hijos y a mi esposa por su apoyo constante para cumplir mis metas.

Agradecimientos

Especial agradecimiento a mi asesor Ing. Dr. Alberto Alva Arévalo Iparraguirre, por su orientación en la ejecución del presente proyecto.

Índice general

Ficha de identificación 7
Dedicatoria 8
Agradecimientos 9
Índice de tablas 11
Índice de figuras 12
RESUMEN 13
ABSTRACT 14
CAPÍTULO I INTRODUCCIÓN A LA INVESTIGACIÓN 15
CAPÍTULO II MARCO TEÓRICO 20
2.1. Antecedentes de la investigación 20
2.2. Fundamentos teóricos 22
2.2.1. Machine learning 22
2.2.2. Sistema De Seguros Vehiculares Contra Accidentes De Tránsito 34
2.2.3. Definición de términos básicos 36
CAPÍTULO III MATERIALES Y MÉTODOS 39
3.1. Ámbito y condiciones de la investigación 39
3.2. Sistema de variables 39
3.3. Procedimientos de la investigación 40
CAPÍTULO IV RESULTADOS Y DISCUSIÓN 53
CONCLUSIONES 63
RECOMENDACIONES 64
REFERENCIAS BIBLIOGRÁFICAS 65
ANEXOS 68

Índice de tablas

Tabla 1. Tipos de Algoritmos de Machine Learning 30
Tabla 2. Información de los campos usados 43
Tabla 3. Evaluaciones de modelos de ML 44
Tabla 4. Fechas de evaluación 53
Tabla 5. Medidas del indicador 1. Afiliaciones 53
Tabla 6. Pruebas de Normalidad del Indicador I 54
Tabla 7. Hipótesis para el indicador – Gestionar el incremento de la cantidad de las afiliaciones con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin 55
Tabla 8. Correlaciones de muestras emparejadas para el indicador I 55
Tabla 9. Prueba de muestras emparejadas para el indicador I 55
Tabla 10. Promedio de costo total de siniestros por años 56
Tabla 11. Fechas de evaluación de estimación de costo de siniestros 57
Tabla 12. Medidas del indicador 2. Predecir gastos de siniestros 57
Tabla 13. Pruebas de normalidad del indicador 2 58
Tabla 14. Hipótesis para el indicador – Pronosticar la Estimación de gastos ocurridos en las coberturas con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin 59
Tabla 15. Correlaciones de muestras emparejadas para el indicador 2 59
Tabla 16. Prueba de muestras emparejadas para indicador 2 60

Índice de figuras

Figura 1. Machine Learning Conceptos y clases……23
Figura 2. Set de Entrenamiento……24
Figura 3. Visualización de tipos de vehículos……25
Figura 4. Tipos de Machine learning……27
Figura 5. ¿Dónde estamos en ML?.....28
Figura 6. Ejemplo de red Asia de Lauritzen y Spegelhalter……30
Figura 7. Entorno de trabajo de azure ML……33
Figura 8. Google Cloud Platform……33
Figura 9. Diseño de la Investigación……40
Figura 10. Modelo de Datos Afiliaciones……41
Figura 11. Modelo de Datos Siniestros……42
Figura 12. Dimensiones del Data Set……43
Figura 13. Pipeline Plot de Decision Tree Classifier……45
Figura 14. Matriz de Confusión……45
Figura 15. Columnas Importantes……46
Figura 16. Curva ROC……47
Figura 17. Reporte de Clasificación……47
Figura 18. DataSet Shape……48
Figura 19. Predicciones del Modelo……48
Figura 20. Modelo de predicción……49
Figura 21. DataSet de Predicciones……49
Figura 22. Muestra Resultados……50
Figura 23. Modelo de machine learning – siniestros……51
Figura 24. Medidas del indicador afiliaciones……54
Figura 25. Pronóstico de costos……57
Figura 26. Pronósticos……58

RESUMEN

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín En 2022

El objetivo general fue determinar la efectividad la implementación del modelo basado en machine learning en la efectividad de la gestión en afiliación y siniestros en AFOCAT San Martin 2022 y los objetivos específicos: Identificar la mejora de las coberturas con la implementación de un modelo basado en machine learning en AFOCAT San Martin y Determinar la mejora de las afiliaciones con la implementación de un modelo basado en machine learning en AFOCAT San Martin. El tipo de la investigación fue Aplicada, el Diseño fue Experimental del Tipo Pre – Experimento, el uso de la Variable dependiente dependerá de los datos proporcionados por la muestra, esta fue de todos los eventos ocurridos en 30 días, la hipótesis general planteo Si a través de la implementación de un modelo basado en machine learning se contribuye significativamente a elevar el nivel de efectividad en la gestión en afiliaciones y siniestros en AFOCAT San Martin, en 2022. La investigación concluyó con que la utilización de un Moledo de machine learning logra gestionar el incremento de las afiliaciones en un 21 % y logra predecir con 96 % de asertividad los gastos de siniestros ocurridos en el 2022.

Palabras clave: Machine Learning, Pronostico, Modelo, Gestión, Aprendizaje Automático, Aprendizaje supervisado, Aprendizaje no supervisado.

ABSTRACT

Machine learning-based model to manage memberships and claims at AFOCAT San Martín during 2022

The general objective was to determine the effectiveness of the implementation of a model based on machine learning in the effectiveness of the management of membership and claims in AFOCAT San Martin 2022. The specific objectives were: To identify the improvement of coverage with the implementation of a model based on machine learning in AFOCAT San Martin and to determine the improvement of membership with the implementation of a model based on machine learning in AFOCAT San Martin. It was an applied research, with an experimental design of the pre-experiment type, the use of the dependent variable depends on the data provided by the sample, which consisted of all events occurring in 30 days. The general hypothesis was that the implementation of a model based on machine learning would contribute significantly to raising the level of effectiveness in the management of affiliations and claims in AFOCAT San Martin, in 2022. The research concluded that the use of a machine learning model manages to increase membership by 21% and is able to predict with 96% accuracy the claims expenses incurred in 2022.

Keywords: Machine Learning, Forecasting, Model, Management, Machine Learning, Supervised Learning, Unsupervised Learning.

CAPÍTULO I
INTRODUCCIÓN A LA INVESTIGACIÓN

El avance tecnológico ocurrido en los recientes años, nos permite acceder a la información en todo tiempo, lugar, dispositivo, medio, idioma, versiones, una gran diferencia desde hace unas décadas atrás, logrando que la sociedad pueda generar datos en volúmenes y cantidades enormes, consumiendo información en dispositivos con excelente procesamiento, en el transcurso de su vida diaria, al escuchar música, podcast, videos, redes sociales y noticias, generando la información necesaria para poder hacer descubrimientos, que nos mejorarán la vida en los siguientes años.

Según Kemp (2024), los datos publicados en el Informe de estadísticas globales digitales en asociación con We Are Social y Hootsuite, podemos observar que 5.34 mil millones de personas son usuarios únicos de un teléfono móvil, así también podemos ver que 5.03 mil millones de personas son usuarios del internet los cuales representan el 63.1 % de la población total, estos datos de acceso están distribuidos en acceso desde distintos dispositivos, podemos comprender también la cantidad de personas conectadas desde distintas zonas del mundo así por ejemplo que el 93 % de personas están conectadas desde el norte del continente americano, en contraste como el 24 % de personas están conectadas desde el centro del continente africano.

De esta información también podemos determinar que YouTube, al ser la segunda web-site más visitada del mundo con más de 14 mil millones de visitantes al mes, esto proporciona aproximadamente 694000 horas de video emitidas por minuto en las mencionada red social, se suben 400 horas de vídeo a YouTube cada minuto en todo el mundo, para poder visualizar todos estos videos tardaremos 82 años, solo para los que se suben en una hora, y para procesar todos los videos almacenados hasta ahora en YouTube nos tomaría alrededor de 180 000 años.

En la misma línea de información proporcionada podemos ver los datos generados por las redes sociales como twitter , Facebook, Instagram, tiktok y alguna otras que puedan surgir, Facebook cuenta con 2.94 mil millones de usuarios activos, siendo el 47.2 % de la población mundial mayores de 13 años que tienen acceso a esta red social, así también Instagram cuenta con 1.44 mil millones de usuarios activos, siendo el 23.1 % de la población mundial mayores de 13 años que tienen acceso a esta red social, siendo las mujeres las que más se conectan a esta red social, en cambio en twitter cuenta con 486.0 millones de usuarios activos, con un porcentaje de participación en internet de 9.7 %.

La gran novedad es Tiktok, esta red social entró con agresividad y con consecuencias en los países que mantienen una animadversión de naturaleza política y económicas, según Arroyo (2022), saliendo al mercado internacional en septiembre de 2017, ha generado un crecimiento tal que cuenta con 1.02 mil millones de usuarios activos mayores de 18 años, generando un tráfico en internet de 20.3%, con un alcance de 18.3 % de la población mundial mayores de 18 años, siendo la red social que proyecta mayor crecimiento y por lo tanto gran cantidad de datos.

Las aplicaciones en donde se usa del machine learning en la actualidad son muy variadas, por ejemplo, en la identificación, por medio de video, de las personas que incumplen las ordenanzas públicas en lo relacionado a la limpieza pública, esto en la ciudad china de Shanghái, controlado por el sistema City Brain, con 290 000 cámaras, es considerado un cerebro, proporcionando al gobierno una forma de dirigir ordenadamente a la ciudad, es un experimento que pronto se realizara en todo el país.

Así también, un ejemplo muy conocido donde se utiliza técnicas de machine learning, es para detectar el correo electrónico que no deseamos como es el spam, en donde por medio de etiquetas y clasificación, los servidores de correo electrónico aprenden a identificar estos correos y remitentes, para colocarlos en listas negras para su posterior difusión en los demás destinatarios de nuestra red de contactos.

Otra industria donde se aplica técnicas de machine learning durante los últimos años es en la generación de texto, imágenes y audio, tanto así que las maquinas pueden aprender a dibujar imágenes solo con texto, tal es el caso del servicio web que ofrece este tipo de procesamiento, Stable Diffusion, las referencias con las que se puede fortalecer al algoritmo con toda la cultura actual es bastante grande, presentándonos una versión para una computadora personal con los requerimientos ideales, así también una API para integración en otros servicios y su respectivo repositorio en GitHub, para colaborar y aprender.

Las técnicas de machine learning también están ayudando y creciendo para poder crear texto, a partir de algunas entradas que configuramos, es el caso de servicios como https://zyro.com/, teniendo opción de generar más texto con opciones de pago.

Otro servicio de IA que esta ayudado a los creadores de contenido para redes digitales, Loudly (2024), que crea música, según los intereses con lo que el algoritmo vaya creciendo hasta obtener lo que nos tenga más satisfecho, así también puede descomponer una canción en varios track por instrumentos, etc.

Lo que causa más admiración en el desarrollo del machine learning es que existen servicios para programar o crear líneas de código, en el lenguaje que necesitemos, ingresando nuestra solicitud en texto, creando una gran variedad de aplicaciones en ingeniería, solucionando las necesidades de las empresas para crecer en el ámbito tecnológico

El Machine learning esta tan cerca de nosotros hoy en día que lo vemos en las recomendaciones de películas en nuestros servicios de streaming favoritos, sea Netflix, Amazon Prime Video u otros, lo vemos en el algoritmo de recomendación de que música necesitamos escuchar proporcionado por spotify, en el los videos de recomendación de YouTube, Facebook y twitter que te recomienda a quien seguir en base a lo que le damos me gusta, compartir o lo reproducimos varias veces, como lo mencione antes el algoritmo de recomendación de tiktok, que nos presenta cada micro video más adictivo que el anterior.

Machine Learning al ser parte de Inteligencia artificial, proporciona alternativas de solución con las cuales podemos predecir las probabilidades de la ocurrencia de una actividad, mediante el aprendizaje de las máquinas de manera supervisada o no supervisada.

Todos estos datos generados por las personas nos proporcionan materia prima esencial para realizar estudios acerca de los distintos comportamientos de las personas, según sus edades, zonas geográficas, horario, sexos, gustos por el contenido de las publicaciones de distintas naturaleza como pueden ser políticos, costumbres, culturales, religiosas, hábitos, etc., generando nuevas métricas y categorías con las cuales podemos realizar nuevos descubrimientos con el fin de ser de ayuda para la sociedad y para la ciencia.

En la presente investigación realizaremos el análisis de estos comportamientos registrados en los distintos sistemas con los que podemos capturar estos datos, como pueden ser, el sistema de ventas, el sistema de gestión de siniestros, sistema de ventas por delivery, ventas en concesionarios, consultas por teléfono, ventas por teléfono, visitas en la paginas web, visitas en las páginas de Facebook, Instagram, etc, para así presentar predicciones para ayudar a la gestión de siniestros y ventas de seguros, con algoritmos de machine learning.

La problemática presentada dentro de AFOCAT SAN MARTIN, ocurre en estos dos procesos principales, como son: las afiliaciones y los siniestros. En lo que denominamos afiliaciones debemos entender que se hace referencia a lo que comercialmente son las

ventas, en este caso venta de seguros, siendo específicos ventas de seguros para vehículos, en las principales categorías y clases, que son determinados por el Ministerio de transportes y comunicaciones.

Las mencionadas afiliaciones se realizan a las personas, que al hacerse acreedor del seguro se convierten asociados, si desean participar de las actividades dentro de la estructura de AFOCAT SAN MARTIN, se pueden convertir en socios y finalmente participar en los procesos ejecutivos como miembros del consejo directivo, que se eligen en asamblea, por un periodo de 2 años.

Uno de los objetivos principales de AFOCAT SAN MARTIN, es como toda empresa, aumentar sus ingresos, para esto generan campañas de ventas, anuncios en redes sociales como Facebook, tiktok, campañas de marketing, regalos por compras, sorteo de premios entre todos los que compran por primera vez un seguro o realizan una renovación de su seguro vehicular, en este contexto se una campaña que genera resultados son las ventas por llamadas telefónicas, cuya data de clientes a los cuales llamar, son aquellos asociados que no renovaron aun su seguro o que esta por vencer en el mes que se realiza el reporte, con un promedio de 2500 llamadas por realizar por ese mes.

Las llamadas para ventas implican un costo de operaciones un poco elevado pues se dedica 2 personas para que realicen estas llamadas, contactarles por mensajes de texto y contacto por mensajería instantánea como WhatsApp, además de otras tareas proporcionadas por la empresa, debido a la naturaleza de su contrato.

Se identifica este proceso como un problema que resolvimos en la presente investigación presentando un modelo de machine learning, que después de analizar la data recolectada, realizamos procesos de extracción, transformación y carga a nuestro cuaderno de trabajo en notebook colaboratory de Google, entrenar el modelo y testearlo, pudimos generar en 21 % en mejorar el indicador de afiliaciones, pues las llamadas para realizar las ventas se redujeron de 2500 a solo 250 aproximadamente como ventas seguras, reduciendo el costo operativo para AFOCAT SAN MARTIN

Otro proceso que se cuenta como principal es la gestión de siniestros, en este caso para la evaluación del posible costo de un siniestro, se gestionan gastos operativos y logísticos, ya que se contrata personal dedicado a realizar esta tarea para cada siniestro reportado y la proyección en los estados financieros es una parte esencial de la supervisión exigida por la Superintendencia de Banca y Seguros en el ámbito de sus funciones regulatorias.

Este es otro problema que se solucionó con la implementación de un modelo de machine learning en AFOCAT SAN MARTIN al lograr predecir los costos de siniestros en 96% los aciertos, generando mejoras en los indicadores de gastos operativos para la mejora financiera de la empresa.

En esta investigación utilizaremos el machine learning para ayudar a la gestión, a las tomas de decisiones, a poder presentar resultados más exactos, donde poder invertir y gestionar adecuadamente los recursos de la institución, mostrar las principales tendencias en clientes, principales gatos de siniestros, estimar los gastos de un siniestrado según sus respectivas características, todo esto a fin de mejorar los servicios y dar calidad de servicio a todos los asociados.

Este informe respecto a la investigación realizada está estructurado de la siguiente estructura: como primer capítulo se consideran a las fuentes bibliográficas sobre antecedentes, en ámbito internacional, ámbito nacional, ámbito de estudio local y el consecuente marco teórico sobre las teorías que sustentan de la investigación, en el segundo capítulo se consignan a aquellos materiales y métodos utilizados durante toda la investigación, en el tercer capítulo se incorporan los resultados obtenidos, así como la discusión que presentan los mencionados resultados. Apartados posteriores, contienen conclusiones, recomendaciones, bibliografía y anexos, donde presentamos los datos técnicos referente a nuestra investigación.

CAPÍTULO II
MARCO TEÓRICO

2.1. Antecedentes de la investigación

Llevando a cabo una revisión de estudios previos realizadas encontramos algunas dignas de poder replicar y sustentar en la presente investigación.

Arévalo (2022), en su tesis de doctorado titulado “Modelo de Gestión Publicitario Basado en Inteligencia Artificial para la Escuela de Posgrado de la Universidad Nacional de San Martín”, el objetivo principal fue evaluar cómo un modelo de gestión publicitaria basado en inteligencia artificial influye en la eficacia de los distintos programas de estudios en la Escuela de Posgrado de la UNSM. Esta investigación se clasificó como aplicada, con un enfoque explicativo. El diseño utilizado en el estudio fue no experimental, aplicada a nivel explicativo con un enfoque cuantitativo. Se seleccionó una muestra de 300 participantes mediante un muestreo aleatorio simple para administrarles una encuesta. Donde concluye que se logró determinar la efectividad del modelo con una (sig. 0,000), se demostró que el modelo fue altamente efectivo y beneficioso en el logro de los objetivos establecidos. Además, se concluyó que el 77% de los profesionales que vieron el anuncio se matricularon en algún programa de maestría.

Aceituno (2019), en su tesis de doctorado titulado “Modelo Predictivo De Análisis De Riesgo Crediticio Usando Machine Learning En Una Entidad Del Sector Microfinanciero”, el objetivo general es identificar el modelo de machine learning más efectivo para mejorar la precisión en la concesión de microcréditos, La investigación fue no experimental de tipo transversal con diseño de investigación descriptiva comparativa, en una muestra de 15015 registros a los cuales se les aplicó distintos tipos de modelos. concluye primeramente que se logró determinar que el modelo de Artificial Neural Network es el más asertivo para los microcréditos, permitiendo reducir el riesgo en el otorgamiento.

Caselli (2021), en su tesis de doctorado titulado “Modelo Predictivo Basado En Machine Learning Como Soporte Para El Seguimiento Académico Del Estudiante Universitario” el objetivo principal de esta investigación fue desarrollar un modelo predictivo basado en Machine Learning para optimizar la gestión del seguimiento académico de los estudiantes universitarios. El diseño de la investigación fue cuasi-experimental, con una muestra seleccionada intencionalmente por el investigador. Esta mencionada investigación concluye que se logra mejorar el seguimiento académico de un 28.89 % a un 58.47 %, así mismo concluye que se logró diseñar un modelo predictivo con Machine

learning y Deep Learning con una secuencia de nodos [56,42,18,14,8,4], finalmente se implementó una red neuronal de perceptrón multicapa donde se logra identificar con una precisión de 81.73 % los alumnos ingresantes con un riesgo de abandono de estudios.

De la Fuente-Carmona (2022), en su tesis doctoral “Diseño de Soluciones Avanzadas Basadas en Técnicas de Machine Learning para la Toma de Decisiones en Gestión de Activos”, el objetivo principal del estudio es desarrollar un proceso que transforme el conocimiento derivado de los datos generados por los activos en herramientas eficaces para la detección de anomalías en dichos activos, para lograr esta investigación desarrolla un proceso para la transformación de conocimiento, desarrolla una herramienta CBM, mediante la revisión de técnicas de machine learning. La conclusión destaca que todo el proceso, que abarca desde el análisis de datos y la extracción de conocimiento mediante la creación de modelos de Machine Learning hasta su aplicación en la gestión del mantenimiento a través de recomendaciones, es fundamental para la transformación digital hacia la industria 4.0.

Bernedo et al. (2021), este grupo de investigadores presentan en su tesis de maestría titulado “Identificación de Obras Urbanas para la Ciudad de Lima a través del uso de Herramientas basadas en Machine Learning”, el objetivo principal de esta investigación es aplicar tecnología basada en machine learning para identificar obras urbanas en la ciudad de Lima, con un enfoque exploratorio. La investigación concluyó que, mediante el modelo de machine learning utilizado, se lograron identificar las zonas de Lima con problemas en las obras públicas, alcanzando una precisión y sensibilidad del 87.4% para parques, del 73.7% para veredas y del 70% para pistas.

Montilla (2022), en su tesis de maestría titulada “Algoritmos de aprendizaje automático supervisado en la predicción del rendimiento académico”, El objetivo principal de este estudio es examinar la efectividad de los algoritmos de aprendizaje automático supervisado en la predicción del rendimiento académico en matemáticas entre los estudiantes de quinto grado de bachillerato de la Institución Educativa "Santa Rosa". Los objetivos específicos son pronosticar el rendimiento académico para el año 2021 utilizando el algoritmo de aprendizaje automático supervisado más efectivo; identificar la relación entre las calificaciones de 2010 a 2020 y la predicción del rendimiento académico para 2021; y proyectar el rendimiento académico en matemáticas de 2021 a 2027. El estudio fundamental, descriptivo, compuesto y predictivo se realizó con una metodología no experimental. Se utilizó una muestra censal de 29.33 alumnos. La hipótesis general afirmaba que existen algoritmos supervisados de aprendizaje automático capaces de predecir el éxito académico de los alumnos de quinto curso de

matemáticas con una precisión del 95% o superior. Los resultados de la investigación confirmaron que, efectivamente, hay algoritmos de aprendizaje automático supervisado que alcanzan o superan una eficiencia del 95% en la predicción del rendimiento académico. Además, se concluyó que la predicción del rendimiento académico en matemáticas para el año 2021 es altamente confiable, La técnica de aprendizaje automático supervisado K-nearest neighbors demostró ser la más eficaz, con una precisión del 100%. No se observó ninguna relación significativa entre las calificaciones de 2010 a 2020 y la previsión de rendimiento académico para 2021 utilizando este enfoque. Asimismo, el pronóstico del rendimiento académico en matemáticas para el período 2021-2027 muestra una tendencia positiva y creciente.

Guzman (2022), en la investigación de maestría titulada “Machine Learning para predecir la adquisición de plataformas educativas de la empresa Difucien Ecuador, 2022”, El objetivo principal era anticipar la adquisición de plataformas educativas. Durante la fase de recogida de datos, se utilizaron fichas de observación tanto para el PreTest como para el PostTest. Tras aplicar la solución Machine Learning, se observaron los siguientes resultados: el tiempo requerido para seleccionar plataformas se redujo en un 18.01% (equivalente a 14 minutos), y el tiempo para identificar clientes potenciales disminuyó en un 32.29% (65.50 minutos). Además, se observó un aumento del 16.50% en el índice de adquisiciones de plataformas. Además, el uso del aprendizaje automático redujo el tiempo necesario para cerrar acuerdos en un 24,35%. En resumen, el uso de un algoritmo de aprendizaje automático mejora drásticamente el proceso de cierre de ventas de artículos académicos.

2.2. Fundamentos teóricos

2.2.1. Machine learning

El Machine Learning es la disciplina que combina ciencia y arte para programar computadoras, permitiéndoles aprender a partir de datos (Géron, 2020). Una definición más orientada a la ingeniería establece que un programa de computadora se considera que aprende a partir de la experiencia “E” en relación con una tarea “T” y una medida de desempeño “P”, si su rendimiento en “T”, evaluado mediante “P”, mejora con la experiencia “E” (Mitchell, 1997).

Según Thompson (2022), En el aprendizaje automático, una computadora examina datos, crea un modelo a partir de esos datos y emplea ese modelo tanto como una hipótesis sobre el entorno como una herramienta de software para solucionar problemas, aquí es donde comenzamos a definir como estudiaremos el machine

learning, aplicado a los datos, iniciando nuestra intención de crear, definir, modelos que permitan demostrar probabilísticamente una suposición y tomarlo como aprendizaje.

El aprendizaje automatizado, conocido como machine learning, se desarrolló como una rama de la inteligencia artificial (IA). Este campo se centra en el uso de algoritmos de aprendizaje que extraen conocimientos de los datos con el fin de generar predicciones (Raschka y Mirjalili, 2021).

Modelos de machine learning

Nos dice Flach (2012), que los modelos de machine learning podemos distinguir los siguientes:

Los modelos geométricos son estructuras en un espacio de instancias con una o más dimensiones. Los datos linealmente separables se definen como aquellos que tienen una frontera de decisión lineal entre clases. Esta frontera de decisión se describe mediante la ecuación (w * x = t), donde (w) es un vector perpendicular a la frontera de decisión, (x) es cualquier punto de la frontera de decisión y (t) es el umbral de decisión.

Los modelos probabilísticos son enfoques diseñados para identificar la distribución de probabilidades que define la relación entre los valores de las características y sus valores asociados. En este contexto, la estadística bayesiana es fundamental, ya que proporciona un marco para actualizar y calcular probabilidades basadas en la evidencia o información previa.

Los modelos lógicos son sistemas que convierten y representan probabilidades en reglas estructuradas mediante árboles de decisión.

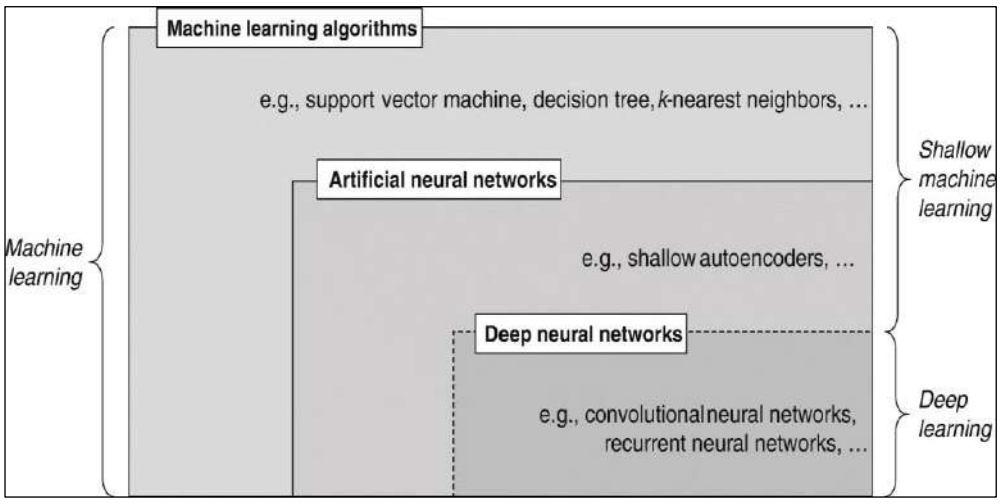


Figura 1
Machine Learning Conceptos y clases
Fuente: *Goodfellow et al. (2016), p. 9*

Dependiendo de la tarea del aprendizaje podemos encontrar otro tipo de distinciones, Figura 1, observamos que los algoritmos de machine learning están basados en redes neuronales artificiales y redes neuronales profundas.

Tipos de Machine learning

Según Raschka y Mirjalili (2021), podemos definir los tipos de machine learning en:

Aprendizaje supervisado

El algoritmo crea una función que representa la conexión entre las entradas del sistema y sus resultados previstos. Un ejemplo común es la clasificación, donde el sistema asigna etiquetas a vectores usando múltiples categorías o clases. La base de conocimientos del sistema se construye a partir de ejemplos previos de etiquetado. Este tipo de aprendizaje tiene aplicaciones significativas en áreas como la investigación biológica, la biología computacional y la bioinformática.

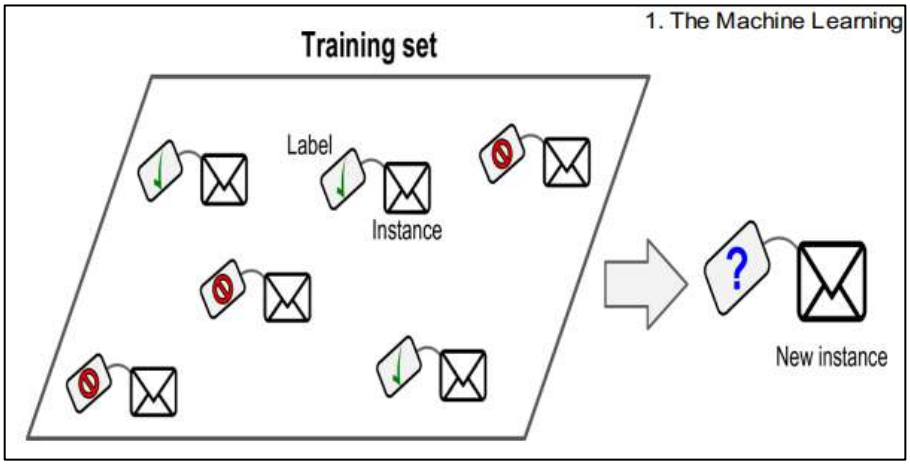


Figura 2
Set de Entrenamiento
Fuente: *Goodfellow et al. (2016)*

En la figura 2, podemos observar que tratamos de entrenar al modelo indicando que ciertos resultados están clasificados y etiquetados, para este caso los mensajes de correo electrónicos clasificados como spam, par cuando realicemos una nueva entrada de datos, este modelo entenderá por las etiquetas o labels, qué tipo de resultado clasificarlo generando un entrenamiento supervisado.

Estos son algunos de los algoritmos de aprendizaje supervisado más importantes: “K-Nearest Neighbors”, “Linear Regression”, “Logistic Regression”, “Support Vector Machines (SVMs)”, “Decision Trees and Random Forests”, “Neural networks” (Capellman, 2020).

Aprendizaje no supervisado

Durante el proceso de modelado, se utiliza un conjunto de ejemplos que incluyen únicamente las entradas al sistema, sin disponer de información sobre las categorías asociadas a estos ejemplos. Por lo tanto, en esta situación, el sistema debe desarrollar la capacidad de identificar patrones para poder asignar etiquetas a nuevas entradas. cómo es de suponer pues no están etiquetados.

A continuación, se enumeran algunos de los algoritmos más destacados en el aprendizaje no supervisado:

- Clustering: K-Means, DBSCAN, Hierarchical Cluster Analysis (HCA).
- Anomaly detection and novelty detection: One-class SVM, Isolation Forest.
- Visualization and dimensionality reduction: Principal Component Analysis (PCA), Kernel PCA, Locally-Linear Embedding (LLE), t-distributed Stochastic Neighbor Embedding (t-SNE).
- Association rule learning: Apriori, Eclat.

Los algoritmos de visualización son ejemplos destacados de algoritmos de aprendizaje no supervisado. Se emplean para analizar volúmenes extensos de datos complejos y no etiquetados, produciendo representaciones en 2D o 3D que facilitan su visualización gráfica.

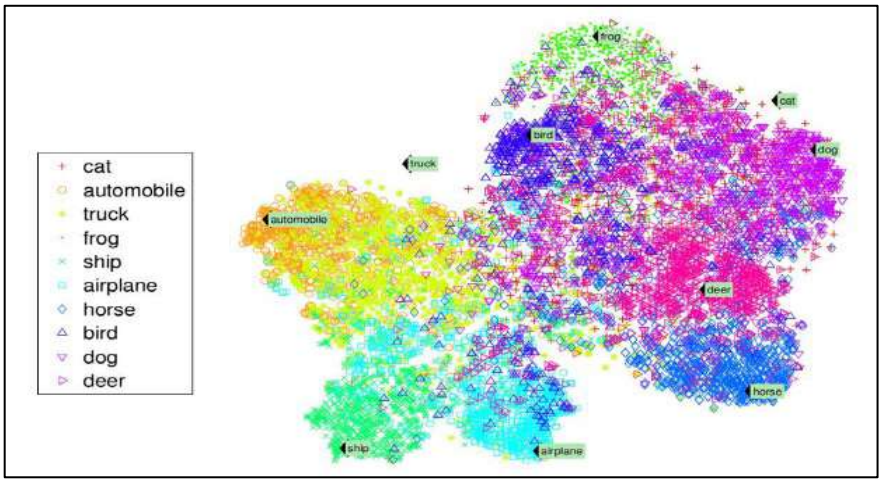


Figura 3
Visualización de tipos de vehículos
Fuente: Capellman (2020, Pag. 26)

Estos algoritmos tratan de preservar la mayor cantidad de estructura posible (por ejemplo, tratando de mantener grupos separados en el espacio de entrada de la superposición en la visualización) con el fin de entender la estructura de los datos y posiblemente descubrir patrones inesperados.

Una tarea afín es la reducción de la dimensionalidad, que busca simplificar los datos minimizando la pérdida de información. Una estrategia para lograrlo consiste en combinar múltiples características correlacionadas en una sola entidad. Por ejemplo, el kilometraje de un automóvil puede estar muy correlacionado con su edad, por lo que el algoritmo de reducción de dimensionalidad los fusionará en una característica que representa el desgaste del coche. Esto se llama extracción de características.

Suele ser una buena idea intentar reducir la dimensión de los datos de entrenamiento mediante una reducción de la dimensionalidad. algoritmo antes de enviarlo a otro algoritmo de aprendizaje automático (como un algoritmo de aprendizaje supervisado). Eso se ejecutará mucho más rápido, los datos ocupan menos espacio en disco y memoria y, en algunos casos, también puede funcionar mejor.

Otra tarea importante no supervisada es la detección de anomalías, por ejemplo, la detección de tarjetas de crédito inusuales. transacciones para evitar el fraude, detectar defectos de fabricación o eliminar automáticamente los valores atípicos de un conjunto de datos antes de alimentarlo a otro algoritmo de aprendizaje. El sistema se muestra principalmente en instancias normales durante el entrenamiento, por lo que aprende a reconocerlos.

Aprendizaje semi supervisado

Estos algoritmos integran tanto el aprendizaje supervisado como el no supervisado para alcanzar una clasificación eficiente. Utilizan tanto datos etiquetados como no etiquetados para mejorar la precisión y la robustez del modelo.

Aprendizaje por refuerzo

El algoritmo aprende detectando su entorno y recibiendo información en respuesta a sus actividades. De este modo, el sistema aprende por ensayo y error.

El aprendizaje por refuerzo constituye la categoría más amplia entre las tres. En lugar de recibir instrucciones directas de un instructor, un agente inteligente debe aprender a interactuar con su entorno a través de recompensas (refuerzos) o castigos, que resultan del éxito o fracaso de sus acciones, respectivamente. El objetivo central es crear una función de valor que ayude al agente a maximizar las recompensas recibidas, mejorando sus estrategias para entender cómo se comporta el entorno y tomar decisiones eficaces para lograr sus objetivos.

Los métodos de aprendizaje por refuerzo más relevantes resuelven problemas de decisión de “Markov” finitos con ecuaciones de “Bellman” y funciones de valor. La

programación dinámica, los enfoques de Montecarlo y el aprendizaje por diferencia de tiempo son técnicas muy conocidas.

“AlphaGo” es un software de inteligencia artificial desarrollado por “Google DeepMind” para participar en el juego de mesa Go. En marzo de 2016, AlphaGo consiguió derrotar a Lee Se-Dol, un jugador profesional de noveno da y ganador de 18 títulos mundiales. Entre los algoritmos que emplea, se destacan el árbol de búsqueda Monte Carlo y el uso de redes neuronales para aprendizaje profundo. Un documental que explora este hito se encuentra disponible en Netflix bajo el título "AlphaGo" (Googledocs, 2023).

Transducción

En contraste con el aprendizaje supervisado, este enfoque no genera explícitamente una función para predecir. En cambio, trata de anticipar las categorías de nuevos ejemplos utilizando los ejemplos de entrada previos, sus respectivas categorías, y los datos nuevos introducidos en el sistema.

Aprendizaje multitarea

Los métodos de aprendizaje emplean el conocimiento previamente adquirido por el sistema para resolver problemas similares a los que ya ha enfrentado. La evaluación del rendimiento y el análisis computacional de estos algoritmos se encuadra en una disciplina estadística llamada teoría computacional del aprendizaje.

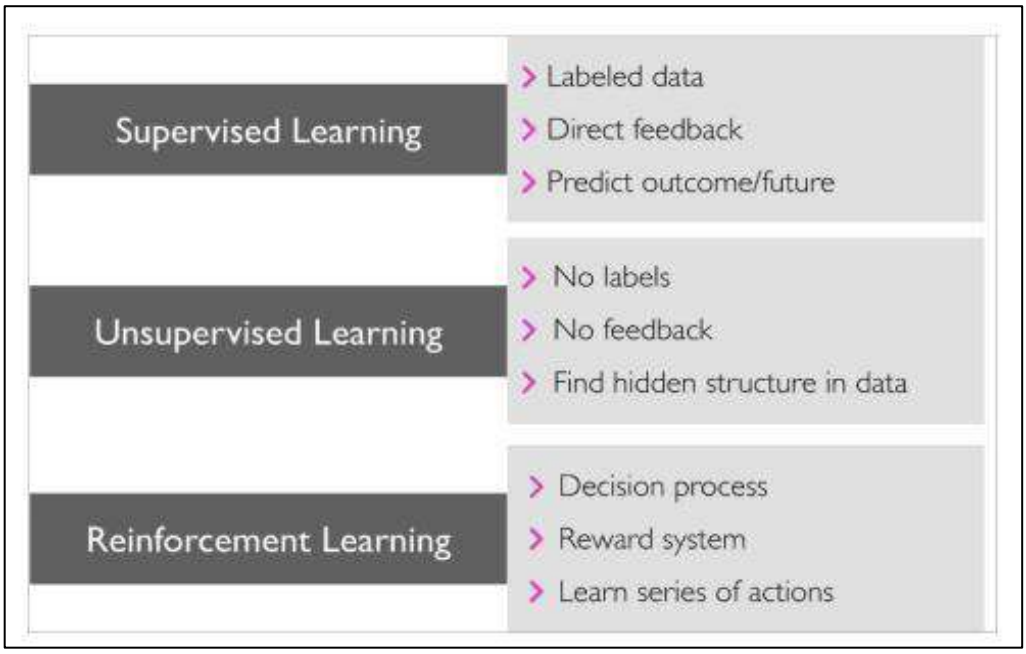


***Figura* 4**
Tipos de Machine Learning
Fuente: Python Machine Learning. Página 35

Técnicas de clasificación

Árboles de decisiones

Este método de aprendizaje utiliza un árbol de decisiones como modelo para realizar predicciones, vinculando observaciones de un objeto con inferencias sobre su resultado final.

Los árboles de elección son diagramas que representan las posibles soluciones a una elección en función de determinados parámetros. Se destacan como uno de los algoritmos más empleados en aprendizaje supervisado dentro de machine learning, capaces de realizar tanto clasificaciones como tareas de regresión (Bagnato, 2022).

Los árboles de decisión empiezan con un nodo inicial denominado raíz, desde donde se ramifican los atributos de entrada en nodos adicionales, cada uno representando una condición que puede evaluarse como verdadera o falsa. Este proceso de bifurcación se repite en cada nodo, subdividiéndose sucesivamente hasta alcanzar las hojas, que son los nodos terminales. Estas hojas proporcionan respuestas definitivas a la solución del problema, como Sí/No, Comprar/Vender, u otras clasificaciones pertinentes.

Para obtener el mejor árbol de decisión posible, se evalúa cada subdivisión entre todas las opciones disponibles, el algoritmo debe evaluar las predicciones obtenidas y compararlas para seleccionar la mejor opción en cada nodo y sus ramas subsiguientes. Para medir y evaluar estas opciones, Se emplean diversas funciones, entre las cuales las más frecuentes y reconocidas son el Índice Gini y la Ganancia de Información, que se fundamenta en la entropía. La división de nodos continuará hasta llegar a la profundidad máxima permitida del árbol o hasta que se cumpla un mínimo establecido de muestras en cada hoja.

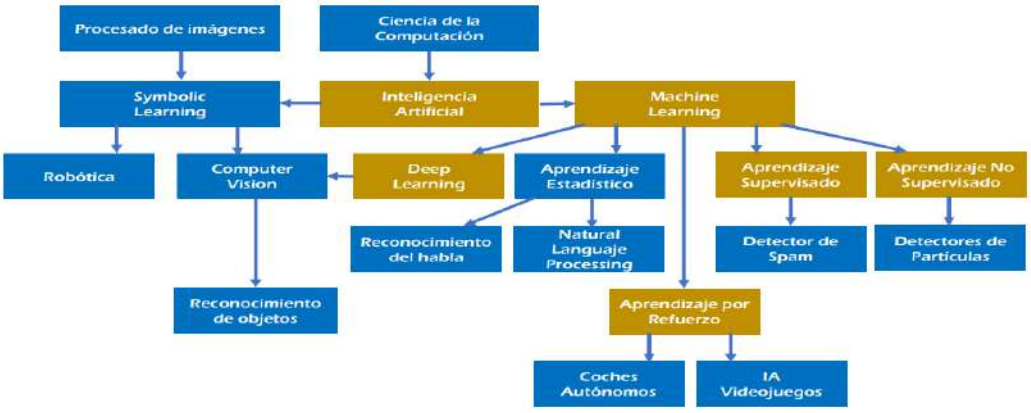


Figura 5
¿Dónde estamos en ML?
Fuente: *© Mg. Joel F. Machado Vicente - UNSM - 2021- Maestría en Ciencias con mención en TI - Machine Learning.*

Reglas de asociación

Los algoritmos de reglas de asociación buscan vínculos significativos entre variables. Los más conocidos son el algoritmo Apriori, Eclat y Frequent Pattern.

Algoritmos genéticos

Los algoritmos genéticos son técnicas de búsqueda basadas en la evolución natural. Emplean operadores como la mutación y el cruce para producir soluciones novedosas y adaptarse eficazmente a problemas únicos.

Redes neuronales artificiales

son un modelo de aprendizaje automático basado en el modo en que se comportan las neuronas en los sistemas nerviosos animales. Estas redes consisten en conexiones entre neuronas que colaboran para generar una salida. Cada conexión tiene un peso numérico que se ajusta según la experiencia, permitiendo así que las redes neuronales aprendan y se adapten a estímulos. Aunque primero perdieron importancia con el avance de los vectores de soporte y los clasificadores lineales, disfrutaron de un resurgimiento a finales de la década de 2000 con la popularización del aprendizaje profundo.

Máquinas de vectores de soporte

Los Métodos de Vectores de Soporte (MVS) son técnicas de aprendizaje supervisado empleadas para clasificación y regresión. Estos algoritmos utilizan un conjunto de ejemplos de entrenamiento que están previamente clasificados en dos categorías, para desarrollar un modelo que pueda predecir si nuevos ejemplos pertenecen a una de estas categorías.

Algoritmos Clustering o agrupamiento

El análisis de agrupamiento, conocido también como clustering, implica la organización de observaciones en subgrupos denominados clusters, donde las observaciones dentro de cada grupo son similares según criterios definidos. Las técnicas de agrupamiento adoptan diferentes enfoques para inferir la estructura de los datos, generalmente basándose en medidas de similitud y niveles de cohesión dentro de cada grupo, así como en la separación entre los distintos grupos. Esta metodología pertenece al ámbito del aprendizaje no supervisado y es ampliamente utilizada en el análisis estadístico de datos.

Redes bayesianas

Una red bayesiana, también conocida como red bayesiana o modelo acíclico dirigido, es un modelo probabilístico que representa un conjunto de variables aleatorias y sus relaciones condicionales mediante un grafo acíclico dirigido. Este tipo de modelo puede representar correlaciones probabilísticas entre variables como enfermedades y síntomas. A partir de datos sobre síntomas concretos, el grafo puede calcular la probabilidad de que aparezcan determinadas enfermedades en una persona. Existen algoritmos eficaces diseñados para inferir y aprender utilizando esta representación probabilística.

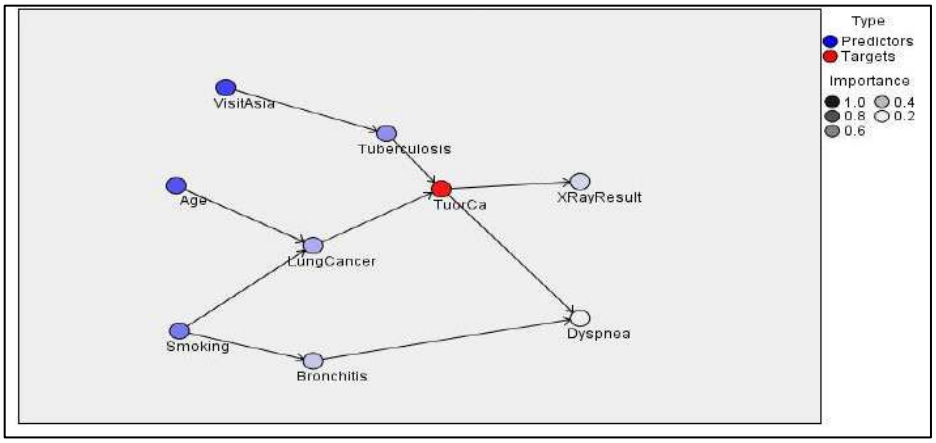


Figura 6

Ejemplo de red Asia de Lauritzen y Spegelhalter

Fuente: © Mg. Joel F. Machado Vicente - UNSM - 2021- Maestría en Ciencias con mención en TI - Machine Learning.

***Tabla* 1**

Tipos de Algoritmos de Machine Learning

ALGORITMOS DE APRENDIZAJE SUPERVISADO			
ítem	Nombre	Característica	Lenguaje
1	Lineal Regression	sklearn linear	Python
2	Polynomial Regression	sklearn	Python
3	Support Vector Regression	StandardScaler	Python
4	Decision Tree Regression	DecisionTreeRegressor	Python
5	Random Forest Regression	RandomForestRegressor	Python
6	Logistic Regression	scikit-learn	Python
7	K-Nearest Neighbors (K-NN)	KNeighborsClassifier	Python
8	Support Vector Machine	scikit-learn	Python
9	Decision Tree Classification	Decision Tree Classifier	Python
10	Random Forest Classification	Random Forest Classifier	Python
ALGORITMOS DE APRENDIZAJE NO SUPERVISADO			
1	K-Means Clustering	Knee Locator	Python
2	Hierarchical Clustering	Agglomerative Clustering	Python

Fuente: Elaboración propia

Deep Learning

Son algoritmos de aprendizaje automático diseñados para capturar abstracciones complejas en los datos, utilizando arquitecturas computacionales que permiten realizar transformaciones no lineales repetitivas de matrices o tensores (Bengio et al., 2013).

Villatoro (2017), menciona que Geoffrey Hinton (2006), El "Deep Learning" se refiere a una rama del aprendizaje automático que emplea algoritmos avanzados para capacitar a las computadoras en la capacidad de "ver" y reconocer objetos, así como discernir texto en imágenes y videos mediante redes neuronales profundas.

No existe una única definición de aprendizaje profundo. En general, se refiere a un tipo de algoritmo construido para el aprendizaje automático. Desde esta perspectiva común, diversas publicaciones se enfocan en características distintas, estos métodos se caracterizan por su estructura de múltiples capas que utilizan unidades de procesamiento que operan de manera no lineal para extraer y modificar características. Cada capa toma la salida de la capa anterior como entrada, permitiendo así la extracción progresiva de características complejas. Estos algoritmos pueden aplicarse tanto en entornos donde se dispone de ejemplos etiquetados para el aprendizaje supervisado como en situaciones donde se aprende de datos no etiquetados en el aprendizaje no supervisado. Sus usos abarcan desde la modelización avanzada de datos hasta la identificación y comprensión de patrones complejos en conjuntos de datos.

Estos enfoques se fundamentan en la adquisición progresiva de características o representaciones de datos a través de varios niveles. Las características más complejas y de alto nivel se construyen a partir de características más simples y de nivel inferior, creando así una estructura jerárquica de representaciones. Este proceso implica aprender múltiples niveles de abstracción que se organizan en una jerarquía, capturando conceptos desde lo más concreto hasta lo más abstracto.

Todas estas descripciones del aprendizaje profundo coinciden en utilizar múltiples niveles de procesamiento que transforman datos de manera no lineal. Estos modelos pueden aprender de forma supervisada o no supervisada para obtener representaciones cada vez más complejas de las características en cada una de sus capas. Estas capas se estructuran jerárquicamente, desde niveles de abstracción más bajos hasta niveles más altos.

Los algoritmos de aprendizaje profundo difieren de los algoritmos de aprendizaje menos profundos en cuanto al procesamiento de la señal desde la entrada hasta la salida. Cada uno de estos procesamientos involucra parámetros entrenables, como pesos y

umbrales. Aunque no existe un consenso general sobre cuántas capas son necesarias para definir un algoritmo profundo, la mayoría de los expertos en la materia creen que el aprendizaje profundo incluye numerosas capas intermedias además de un número mínimo.

Microsoft Azure Machine Learning

Azure Machine Learning capacita a científicos de datos y desarrolladores para crear, desplegar y gestionar modelos avanzados de manera eficiente y segura. Acelera el tiempo necesario para obtener resultados valiosos mediante MLOps, esta plataforma está configurada para ofrecer aplicaciones de inteligencia artificial responsable mediante aprendizaje automático, asegurando la interoperabilidad con código abierto y herramientas integradas., asegurando la seguridad y la confianza en su uso (Microsoft, 2023).

sus principales características ofrecidas son:

- Desarrollo ágil y eficiente de modelos utilizando herramientas integradas y soporte total para bibliotecas y frameworks de código abierto.
- Creación de modelos de inteligencia artificial con enfoque en la responsabilidad, asegurando equidad y transparencia, además de cumplir con estándares éticos.
- Implementación, gestión y colaboración efectiva en el despliegue de modelos de machine learning entre diferentes áreas y operaciones de machine learning.
- Integración completa de gobernanza, seguridad y cumplimiento para ejecutar cargas de trabajo de aprendizaje automático en diversos entornos de manera segura.

Funcionalidades de servicio que ofrece Azure:

Etiquetado de datos, preparación de datos, cuadernos colaborativos, aprendizaje automático automatizado, aprendizaje automático de arrastrar y soltar, Aprendizaje de refuerzo, compilación responsable, experimentación, registros, integración con Git y GitHub, Puntos de conexión administrados, Proceso de autoescalado, Interoperabilidad con otros servicios de Azure, integración con Power BI y servicios como Azure Synapse Analytics, Azure Cognitive Search, Azure Data Factory, Azure Data Lake, Azure Arc, Azure Security Center y Azure Databricks, Compatibilidad con entornos híbridos y multinube, Seguridad de nivel empresarial, Cost Management.

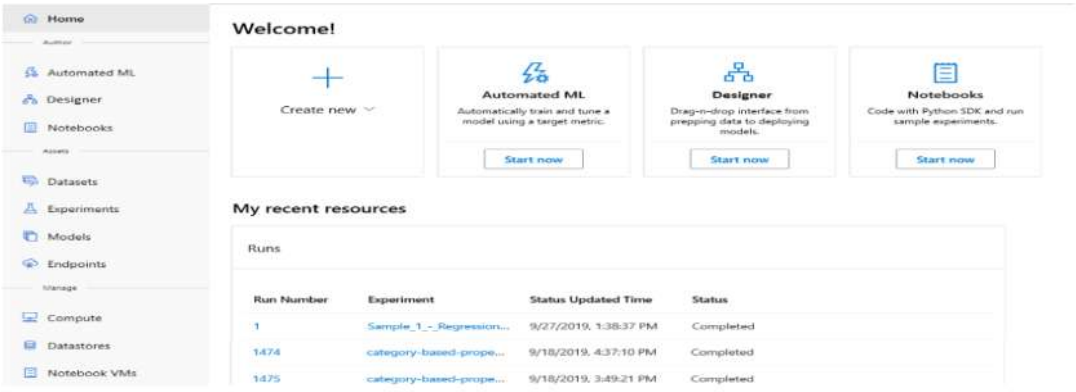


Figura 7
Entorno de trabajo de azure ML
Fuente: https://azure.microsoft.com/es-es/products/machine-learning/#security

Google Cloud Platform

Google Cloud incluye tanto recursos físicos, como ordenadores y discos duros, como recursos virtuales, como máquinas virtuales (VM), que se alojan en los centros de datos de Google repartidos por todo el mundo. Cada centro de datos se encuentra en una ubicación distinta, como Asia, Australia, Europa, Norteamérica y Sudamérica. Cada área consta de zonas diferenciadas, proporcionando redundancia y alta disponibilidad dentro de la misma región.

En la computación en la nube, los productos tradicionales de software y hardware se transforman en servicios accesibles, permitiendo el uso de recursos subyacentes de manera eficiente. La lista de servicios disponibles en Google Cloud es extensa y continúa expandiéndose. Al desarrollar tu sitio web o aplicación en Google Cloud, integras y combinas estos servicios para construir la infraestructura necesaria. Luego, agregas tu código para crear el producto final según tus requisitos específicos.

Figura 8
Google Cloud Platform
Fuente: https://cloud.google.com/docs/overview

Amazon Web Services

AWS es una plataforma integral de servicios de computación en la nube pública proporcionada por Amazon.com a través de Internet. Estos servicios, también conocidos como servicios web, Dropbox, Foursquare y HootSuite son ejemplos de aplicaciones populares que las emplean. AWS es líder mundial en computación en la nube, compitiendo directamente con servicios como Microsoft Azure, Google Cloud Platform e IBM Cloud. AWS, un reconocido pionero de la industria, proporciona un conjunto diverso de herramientas y recursos para ayudar con la creación y ejecución de aplicaciones basadas en la nube.

Amazon Elastic Compute Cloud (Amazon EC2) es un componente integral de Amazon Web Services (AWS), la infraestructura de computación en la nube de Amazon.com. EC2 permite a los clientes alquilar máquinas virtuales desde las que pueden ejecutar sus aplicaciones de forma flexible y escalable. Este servicio ofrece un cambio en el negocio de la informática al proporcionar potencia de cálculo que se adapta a la demanda, con un pago basado únicamente en el tiempo y los recursos empleados. En lugar de adquirir o arrendar hardware específico a largo plazo, en EC2 se alquila capacidad de computación por el tiempo que sea necesario, permitiendo una mayor flexibilidad y eficiencia en costos para los usuarios.

Amazon Cloud Drive es un servicio de almacenamiento de archivos que Amazon lanzó el 29 de marzo de 2011. Ofrece a los clientes 10 gigabytes de espacio de almacenamiento gratuito, con la opción de adquirir gigabytes adicionales a un costo de un dólar por cada gigabyte por año.

Junto con Amazon Cloud Drive, está disponible un servicio de música llamado Cloud Player, que permite a los clientes acceder a su música desde cualquier ordenador o dispositivo Android conectado a Internet. Sin embargo, se ha puesto en duda la legalidad de Cloud Player, ya que Amazon no compra los derechos de autor de la música antes de distribuir la aplicación. A pesar de ello, Amazon decidió ofrecer este servicio para competir en el mercado y atraer la atención de los consumidores.

Cloud Player, el servicio musical de Amazon, solo está disponible para usuarios en Estados Unidos. Aunque los archivos pueden registrarse y cargarse desde cualquier lugar, el acceso a la música en la nube está prohibido fuera de Estados Unidos.

2.2.2. Sistema De Seguros Vehiculares Contra Accidentes De Tránsito

Según el marco legal del SOAT y CAT, presentamos las principales normativas.

Ley General de Transporte y Tránsito Terrestre, Ley Nº 27181.

El Reglamento para la Supervisión de las Asociaciones de Fondos Regionales o Provinciales Contra Accidentes de Tránsito (AFOCAT) y para el Funcionamiento de la Central de Riesgos de Accidentes de Tránsito, aprobado por Decreto Supremo Nº 040-2006-MTC y sus modificatorias, describe las normas y procedimientos que rigen a las asociaciones de fondos regionales o provinciales para el pago de accidentes de tránsito (Reglamento AFOCAT) (Ministerio de Transportes y comunicaciones [MTC] (2024).

Resolución Ministerial Nº 1075-2016-MTC/01.02, que aprueba el Certificado contra Accidentes de Tránsito (CAT) calcomanía y holograma de seguridad.

Texto Único Ordenado del Reglamento de Responsabilidad Civil y Seguros Obligatorios de Accidentes de Tránsito, aprobado por Decreto Supremo N° 024-2002-MTC y sus modificatorias (Reglamento SOAT).

Reglamento del Fondo de Compensación del Seguro Obligatorio de Accidentes de Tránsito (SOAT) y del Certificado contra Accidentes de Tránsito (CAT), aprobado por Decreto Supremo N° 024-2004-MTC y sus modificatorias.

Ley General del Sistema Financiero y del Sistema de Seguros y Orgánica de la Superintendencia de Banca y Seguros, Ley Nº 26702 y sus modificatorias.

Podemos identificar 2 tipos de seguros, en estos casos, los de uso nacional (SOAT) y los de cobertura regional (AFOCAT) con su Certificados contra accidentes de tránsito.

Tipos de coberturas: Según el Artículo 32 del Reglamento AFOCAT, Decreto Supremo Nº 040-2006-MTC, definimos 5 tipos de coberturas.

- Muerte, cobertura por cada uno de los fallecidos en el accidente, con Cuatro (4) UIT.

- Invalidez permanente, cobertura por cada uno de los fallecidos en el accidente hasta Cuatro (4) UIT.

- Incapacidad temporal, cobertura por cada uno de los fallecidos en el accidente hasta una (1) UIT.

- Gastos médicos, cobertura por cada uno de los fallecidos en el accidente hasta Cinco (5) UIT.

- Gastos de sepelio, cobertura por cada uno de los fallecidos en el accidente hasta: Una (1) UIT.

2.2.3. Definición de términos básicos

- TensorFlow: Es una herramienta desarrollada por Google que es compatible con varios idiomas y plataformas, y está distribuida bajo la licencia Apache 2.

- Apache Mahout: Es una plataforma basada en Java que se enfoca en algoritmos de aprendizaje automático que pueden escalarse eficientemente, especialmente en áreas como el filtrado colaborativo, agrupamiento (clustering) y clasificación.

- Dlib: Una biblioteca bajo licencia Boost para desarrollar en C++.

- KDD: es una metodología, completo y sistemático que se utiliza para descubrir información valiosa y conocimiento a partir de grandes cantidades de datos

- ELKI: Es una plataforma para Java con licencia AGPLv3.

- Encog: es un framework de aprendizaje automático que está disponible para Java y .NET. Ofrece soporte para una variedad de algoritmos de aprendizaje, incluyendo redes bayesianas, modelos ocultos de Markov y máquinas de vectores de soporte. Sin embargo, destaca principalmente en el desarrollo y aplicación de algoritmos de redes neuronales.

- KNIME: Es una herramienta de minería de datos que facilita la creación de modelos utilizando una interfaz visual. Está integrada en la plataforma Eclipse para ofrecer un entorno de desarrollo familiar y robusto.

- Mlpy: Es una biblioteca de aprendizaje automático de código abierto de Python construida sobre NumPy/SciPy, la biblioteca científica de GNU y hace un uso extensivo del lenguaje

- Cython: es un lenguaje de programación diseñado para facilitar la creación de extensiones para Python utilizando C y C++. Aunque su sintaxis es similar a la de Python, Cython permite integrar funciones en C o C++ directamente en el código, lo que amplía significativamente las capacidades de Python en términos de rendimiento y funcionalidad.

- OpenCV: Es una biblioteca de código abierto para el desarrollo de aplicaciones de visión artificial, inicialmente creada por Intel

- OpenCV significa Open Computer Vision

- R: Es un lenguaje de programación enfocado en estadísticas que cuenta con diversas bibliotecas especializadas en aprendizaje automático, como e1071, rpart, nnet, randomForest, entre otras.

- RapidMiner: Es una herramienta de software diseñada para analizar y extraer información de datos. Facilita el desarrollo de procesos analíticos al permitir la conexión visual de operaciones consecutivas. Es utilizada en diversos campos como investigación, educación, capacitación, prototipado rápido y aplicaciones empresariales.

- scikit-learn: Biblioteca en Python que interactúa con NumPy y SciPy

- Spark MLlib: Es una biblioteca integrada en Apache Spark, una plataforma diseñada para realizar cómputo distribuido a gran escala.

- Python: Es un lenguaje de programación fácil de leer y entender, ampliamente utilizado en el desarrollo de aplicaciones diversas como Instagram, Netflix, Spotify, Panda 3D y otros proyectos destacados.

- Jupyter: El Proyecto Jupyter es una entidad sin fines de lucro dedicada a la creación de software de código abierto, estándares abiertos y servicios que facilitan la computación interactiva en múltiples lenguajes de programación.

- Google Colab: Colaboratory, conocido también como "Colab", es una herramienta desarrollada por Google Research que permite a los usuarios escribir y ejecutar código Python directamente desde el navegador. Está diseñado especialmente para actividades como aprendizaje automático, análisis de datos y propósitos educativos.

- Anaconda: Anaconda es una plataforma de distribución gratuita y de código abierto que incluye los lenguajes Python y R, orientada principalmente a la ciencia de datos y el aprendizaje automático. Permite realizar análisis de grandes volúmenes de datos, predicciones analíticas y cómputo científico.

- CAT: Certificado contra Accidentes de tránsito, autorizado para activar las coberturas que las AFOCAT ofrecen.

- Tomadores de CAT: Individuos o entidades que adquieren el Seguro Obligatorio contra Accidentes de Tránsito (CAT) para estar protegidos en caso de sufrir accidentes viales.

- Matricial: En una estructura matricial, también conocida como sistema de mandos múltiples, una organización opera con dos tipos de estructuras al mismo tiempo.

Esto significa que los empleados reportan a dos jefes diferentes y trabajan dentro de dos líneas de autoridad distintas.

- Tensorial: Un tensor es una entidad algebraica que especifica una relación multilineal entre conjuntos de objetos matemáticos dentro de un espacio vectorial. Puede mapear vectores, escalares y otros tensores, estableciendo conexiones y operaciones en contextos matemáticos y físicos.

- Gestión: Son actividades ejecutadas de manera efectiva y apropiada con el propósito de alcanzar objetivos específicos, generando resultados que contribuyen a resolver problemas y alcanzar metas establecidas.

- Afiliaciones: Es aquella persona u organización social, que decide inscribirse a la asociación del fondo contra accidentes de tránsito.

- Siniestros: Es un evento que causa daño significativo o pérdida material. También se refiere al evento específico cubierto por un contrato de seguro, que activa la responsabilidad del asegurador para proporcionar la cobertura acordada.

- Accidentes de tránsito: Se denomina también como incidente de tráfico, siniestro automovilístico, o accidente de carretera, entre otras denominaciones. Se trata de un evento que usualmente ocurre cuando un vehículo choca contra algún elemento de la vía, como otro vehículo, un peatón, un animal, escombros en el camino, así como estructuras estacionarias como postes, edificios o árboles.

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1. Ámbito y condiciones de la investigación

3.1.1 Contexto de la investigación

Se toma como ubicación política y geográfica, la región San Martin, tomando en cuenta los lugares donde AFOCAT SAN MARTIN, cuenta con puntos de venta y el lugar donde sucedan los accidentes de tránsito, todo esto bajo la normatividad vigente, según Texto Único Ordenado del Reglamento de Responsabilidad Civil y Seguros Obligatorios de Accidentes de Tránsito, aprobado por Decreto Supremo N° 024-2002-MTC y sus modificatorias (Reglamento SOAT).

3.1.2 Periodo de ejecución

Esta investigación se realizó por el plazo de 3 meses, tomando como inicio el 01 de agosto de 2022 y como fin del proceso el 30 de octubre de 2022, coincidiendo con el cronograma de actividades.

3.1.3 Autorizaciones y permisos

No aplica

3.1.4 Control ambiental y protocolos de bioseguridad

No aplica

3.1.5 Aplicación de principios éticos internacionales

En este acápite cumplimos con los principios éticos y reglamentos proporcionados por la Universidad Nacional de San Martin, con Resolución N° 1312-2021-UNSM/CU-R, donde se aprueba el reglamento general de investigación, y bajo el código de ética para la investigación científica aprobada con Resolución N° 556-2022-UNSM/CU-R.

La investigación basa como estrategia para la propiedad intelectual Resolución N° 1099-2018-UNSM/CU-R/NLU y según el manual de investigación DIRECTIVA No 001-2022-UNSM/VRINV

3.2. Sistema de variables

3.2.1.1 Variables principales

Variable dependiente: “Nivel de Efectividad en afiliaciones y siniestros en AFOCAT San Martin 2022”

Variable independiente: “Modelo basado en Machine Learning”

Variables secundarias

No aplica

3.3. Procedimientos de la investigación

Entre los materiales utilizados para el desarrollo de los objetivos se considera, google sheets, notebook google colaboratory, jupyter Notbooks.

Se considera para la muestra de nuestra investigación todos los registros encontrados desde el inicio de actividades en el 2014 hasta el 2021

La técnica de investigación se utilizó a nivel explicativo. Se adoptó un diseño experimental preexperimental, con la variable dependiente derivada de los datos de la muestra.

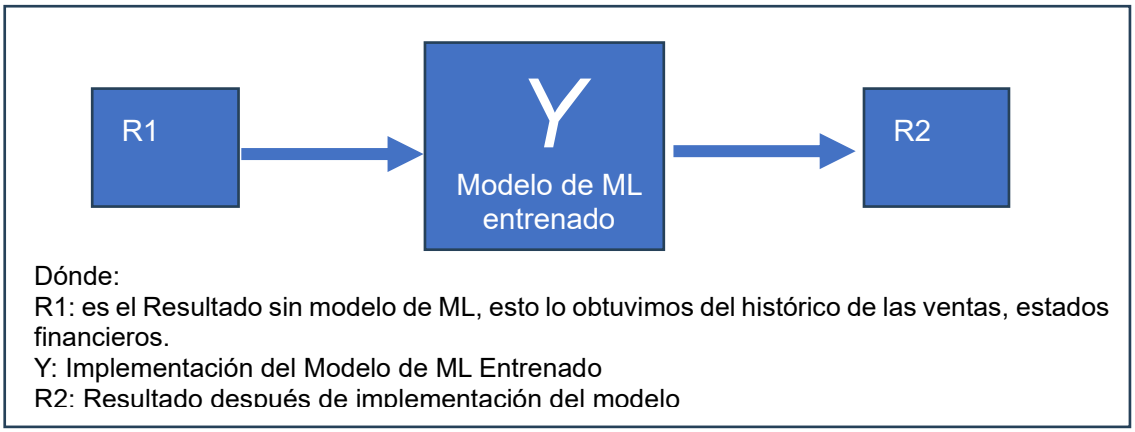


Figura 9
Diseño de la Investigación
Fuente: Elaboración Propia

Como parte del procedimiento de investigación definiremos los puntos necesarios para la caracterización de nuestro modelo de machine learning a partir de la data encontrada en los distintos sistemas informáticos y medios de almacenamiento de estos.

Selección de Características: Identificaremos las variables relevantes que pueden influir en la predicción o clasificación del modelo, para la gestión de afiliaciones y la gestión de siniestros, tomaremos los siguientes esquemas de datos para la construcción de nuestro modelo:

Encontramos que la data para nuestro modelo de afiliaciones, tenemos las entidades de la base de datos del sistema de ventas de AFOCAT SAN MARTIN y de su diccionario

de datos, bd_afocat_sm, son las siguientes: personas, asociados, vehículos, ventas, certificado, categoría, clase, tipo vehículo, tipo servicio, tipo persona, detalle ventas, etc.

Figura 10
Modelo de Datos Afiliaciones
Fuente: Base de Datos Afocat San Martin

Realizamos los mismos procedimientos para el objetivo de gestión de siniestros.En el diccionario de datos de la base de datos del sistema de siniestros de AFOCAT SAN MARTIN encontramos alguna de las siguientes entidades del modelo de datos: eventos, siniestros, accidentados, nosocomios, vehículo, tipo cobertura, tipo accidente, tipo carretera, adjuntos, pagos, pagos pendientes, procedencia, comisaria, tipo nosocomio, pagos vencidos, uit, etc.

Figura 11
Modelo de Datos Siniestros
Fuente: Base de Datos Afocat San Martin

Extracción de Características: de nuestros esquemas realizados, extraemos los campos relevantes hasta da con aquellas entidades, campos u registros calculados que se adecuen a nuestro modelo, Se realizo algunos pasos como son:

Eliminación de las características de cardinalidad alta o sin variación (por ejemplo: hashes, id, imágenes, links, DNI, Nombres y Apellidos, Direcciones, teléfonos o datos considerados sensibles o GUID), Atribución de los valores que faltan, para el caso de columnas que presenten datos nulos o en blanco, que podrían generar dispersión en la normalidad de la data, así como son los datos que generan error estadístico disperso, todo a conveniencia del investigador.

Generación de más características, como el cálculo de formatos de hora en 24 horas, separación de fechas en años, meses, días, cálculo de edades, entre otros.

Quedando nuestro data set de trabajo como sigue:

Tabla 2

Información de los campos usados

item	Comuna	Non	“Null Count”	“Dtype”
0	Sexo	1598584	“non-null”	“int64”
1	dia_nac	1598584	“non-null”	“int64”
2	mes_nac	1598584	“non-null”	“int64”
3	anno_nac	1598584	“non-null”	“int64”
4	Edad	1598584	“non-null”	“int64”
5	Tipo_Persona	1598584	“non-null”	“int64”
6	ID_MARCA	1598584	“non-null”	“int64”
7	Modelo	1598584	“non-null”	“int64”
8	anno_v	1598584	“non-null”	“int64”
9	Nro_Asientos	1598584	“non-null”	“int64”
10	Categoria_Vehiculo	1598584	“non-null”	“int64”
11	Uso_vehiculo	1598584	“non-null”	“int64”
12	Clase_Vehicular	1598584	“non-null”	“int64”
13	Tipo_Carroceria	1598584	“non-null”	“int64”
14	Fecha_Emision_Cat	1598584	“non-null”	“Object”
15	dia_emision	1598584	“non-null”	“int64”
16	mes_emision	1598584	“non-null”	“int64”
17	anno_emision	1598584	“non-null”	“int64”
18	Hora_Emision	1598584	“non-null”	“int64”
19	min_emision	1598584	“non-null”	“int64”
20	cNRD	1598584	“non-null”	“int64”
21	Ambito_Aplicacion	1598584	“non-null”	“int64”
22	Placa_V	1598584	“non-null”	“Object”
23	Punto_Venta	1598584	“non-null”	“int64”
24	lejania	1598584	“non-null”	“int64”
25	nveces	1598584	“non-nul”	“int64”
26	motivo	1598584	“non-un”	“int64”
27	tipo_socio	1598584	“non-null”	“int64”

Fuente: Elaboración Propia

Se debe mencionar que el diccionario de datos de nuestro data set está definido en los anexos de la presente investigación.

Figura 12

Dimensiones del Data Set

Fuente: Elaboración Propia

En la Figura anterior podemos observar que el tamaño de nuestro data set. Que consta de 1 598 584 registros y 28 columnas

3.3.1 Objetivo específico 1: Gestionar el Incremento de la cantidad de las afiliaciones con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin

Para lograr este objetivo se desarrollaron las siguientes actividades:

1. Recopilación de datos de los reportes de ventas, afiliaciones, reportes a la SBS, estados financieros, publicaciones y archivos, tanto digitales y físicos, que se tienen desde que la AFOCAT inicio sus operaciones, desde el 2008 hasta el 2022, que es donde se efectuó la investigación.

2. Limpieza y preparación de datos usando metodología KDD, que la poseer data de distintas fuentes, como archivos xls, txt, vsc, pdf, img y back-up de motores de base de datos como Oracle, Sql Server 2008 R2, MySql Server y MS Acces.

3. Selección del modelo de Machine Learning que según los datos y objetivo que queremos lograr determinamos el algoritmo de aprendizaje automático supervisado como modelo que ejecutamos.

De nuestra evaluación de los conjuntos de datos designamos nuestra variable principal para el modelo de machine learning en este caso será cNRD, que indica su un cliente se afilio como nuevo o renovación a su respectiva afiliación.

***Tabla* 3**

Evaluaciones de modelos de ML

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	0.9254	1.0000	1.0000	1.0000	10.505
rf	Random Forest Classifier	0.8015	1.0000	0.9954	0.8015	1.0000	1.0000	1.0000	117.03
lr	Logistic Regression	0.9718	0.9932	0.9854	0.984	0.9847	0.8015	0.8016	121.768

CPU times: User 8min 56s, sys: 1min 11s, total: 10min 7s Wall time: 49min 34s

Fuente: Notebook Google Colaboratory – Elaboración Propia

Donde podemos observar lo siguiente respecto a la evaluación de los modelos, utilizando nuestro data set invocando a la librería pycaret:

- Recall: Es una medida de qué tan bien el modelo puede identificar todos los casos positivos. La fórmula utilizada seria Recall = TP / (TP + FN)
- Precision: Mide la capacidad del modelo para identificar solo los casos positivos, sin identificar incorrectamente algunos casos negativos, cuya expresión para calcularla seria Precision = TP / (TP + FP)

- F1 Score: Mide la capacidad del modelo para identificar correctamente los casos positivos y evitar identificar incorrectamente los casos negativos. Esto lo calculamos aplicando la siguiente configuración

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Evaluamos que para nuestro modelo obtenemos la mejor precisión (Accuracy) ver tabla 3, en este caso **Decision Tree Classifier**, que es una técnica de aprendizaje supervisado que se puede utilizar tanto para problemas de **clasificación** como de regresión.

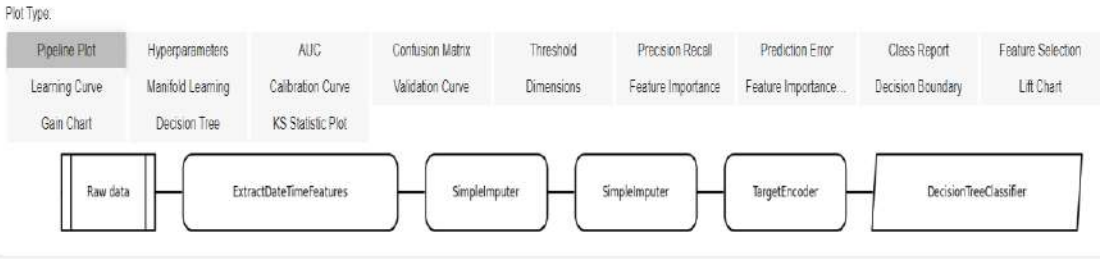


Figura 13
Pipeline Plot de Decision Tree Classifier
Fuente: Pycarest – Python

El modelo toma los datos en bruto proporcionado por nuestro dataset, Raw data, luego ejecuta la extracción de datos de tiempo, identifica los tipos de datos tiene cada columna o variable a evaluar, con SimpleImputer vuelve a ingresar los datos para el árbol de decisión y ejecutar su clasificación para obtener la mejor puntuación, presentando los datos para nuestro pronostico.

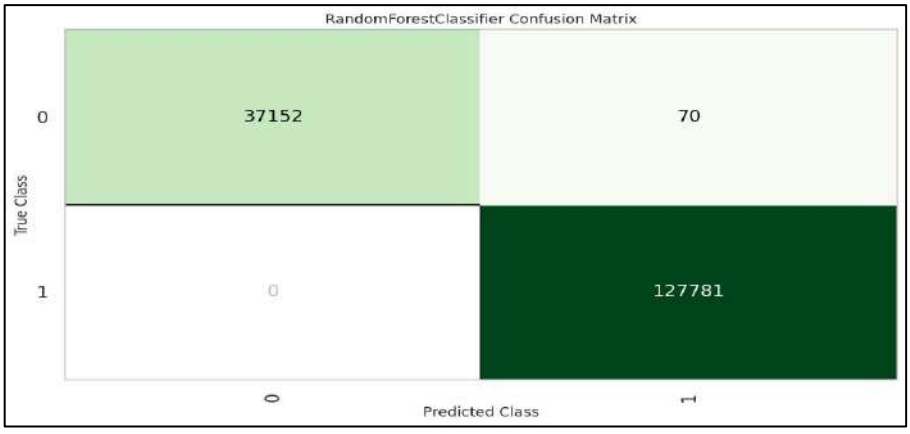


Figura 14
Matriz de Confusión
Fuente: Pycarest – Python

Podemos ver que nuestro modelo predice que para nuestra variable de estudio cNRD, hubo 127781 aciertos en renovación de afiliaciones y 37125 para nuevas afiliaciones.

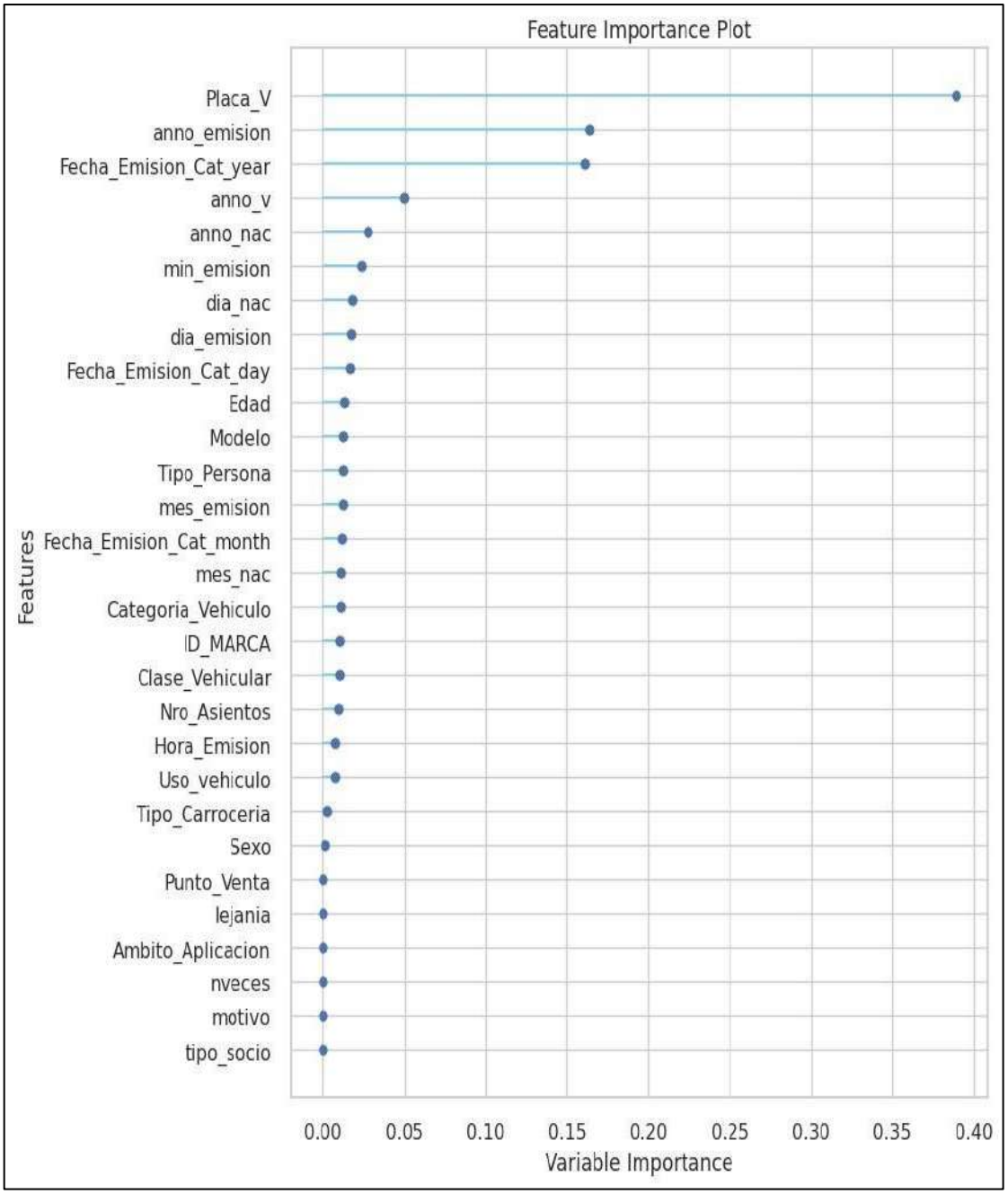


Figura 15
Columnas Importantes
Fuente: Pycarest – Python

Podemos observar en el grafico anterior que para nuestro modelo de las columnas Placa, año de emisión y año del vehículo son las variables que considera de importancia, por las veces que aparece en nuestro conjunto de datos e implica que el cliente va a renovar su afiliación o realizara una compra por primera vez.

Así también podemos ver que las columnas sexo, punto de venta, lejanía, ámbito de aplicación, veces, motivo y tipo socio no son relevantes para el modelo pues para el árbol de clasificación con este data set no tienen ninguna implicancia.

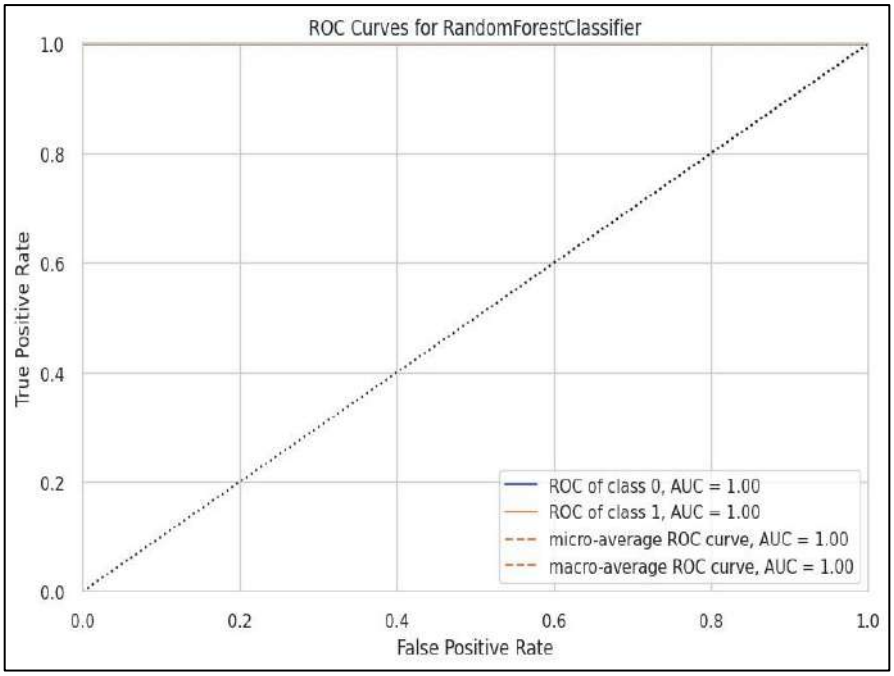


Figura 16
Curva ROC
Fuente: Pycarest – Python

Observamos que lo datos están por encima de nuestra curva ROC, lo cual indica que los modelos aceptan los datos proporcionados ya que no existen falsos positivos, proporcionando acertadas respuestas a nuestro modelo.

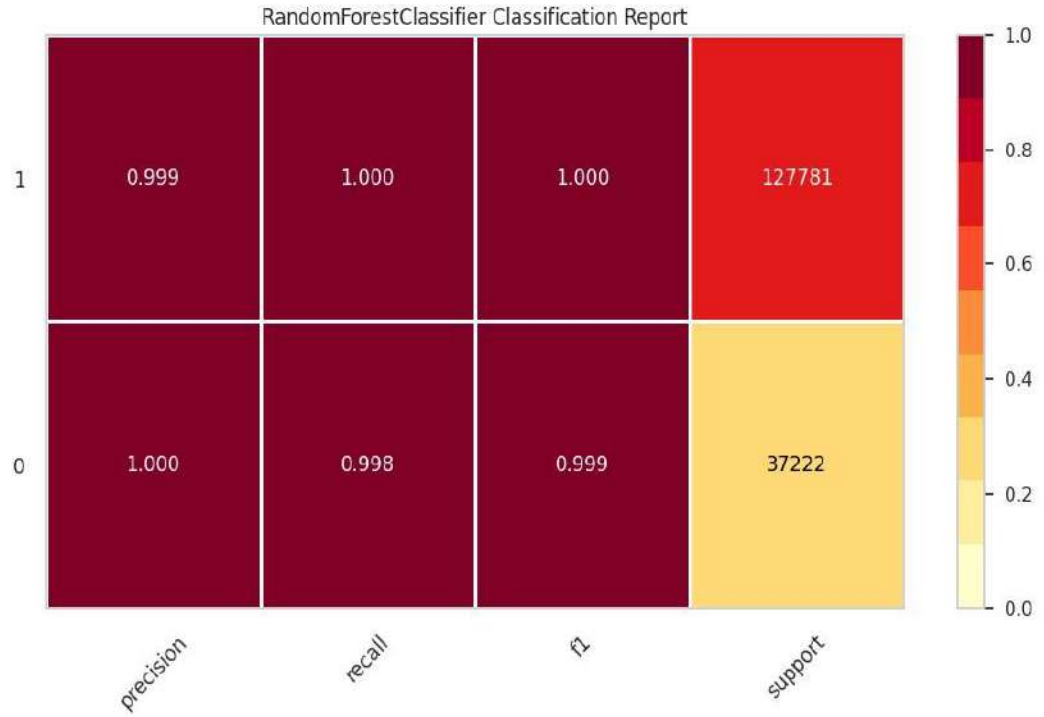


***Figura* 17**
Reporte de Clasificación
Fuente: Pycarest – Python

Observaos la precisión que se tiene para cada uno de los clientes que renovaron o realizaron una compra nueva de su Cat, acertando la mayoría de los casos.

4. Entrenamiento del modelo y Validación del modelo realizado para la investigación, esto lo realizamos en el entorno de desarrollo Anaconda en cuadernos de trabajo de jupyter notebook con el lenguaje de programación Python en su versión 3.11.4

Visualizamos la cantidad de filas y columnas de nuestro data set:

```
[ ] df.shape

    (1598584, 28)
```

Figura 18
DataSet Shape
Fuente: Pycarest – Python

Aplicamos a nuestro dataset, nuestro modelo predictivo, respectivamente evaluado, en las descripciones anteriores, resultando en lo siguiente:

```
[44] predict_model(mod)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Decision Tree Classifier	0.9950	0.9789	0.9979	0.9966	0.9973	0.9647	0.9647

	Sexo	dia_nac	mes_nac	anno_nac	Edad	Tipo_Persona	ID_MARCA	Modelo	anno_v	Nro_Asientos	...	Ambito_Aplicacion	Placa_V	Punto_Venta	lejania	nveces	motivo	tipo_socio	cNRO	prediction_label	prediction_score
48029	1	13	9	1965	56	1	1	41	2006	2	...	9	MX-37841	1	2	5	5	1	1	1	1.00
1567420	1	31	12	2002	18	2	2	4	1997	5	...	9	M3D-028	31	4	5	1	2	1	1	1.00
1436513	1	21	6	1957	64	1	2	4	1997	5	...	9	M3D-554	1	4	1	1	3	1	1	1.00
1351960	1	7	1	1974	47	1	6	44	2017	7	...	6	M4J-203	1	3	2	5	2	1	1	1.00
378382	1	1	1	1980	41	1	2	6	2011	5	...	9	B9E-234	19	1	1	1	3	1	1	1.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1307753	1	12	4	1982	39	1	2	70	2020	5	...	9	S1N-635	1	2	2	1	1	0	0	1.00
1347252	1	16	9	1962	59	1	2	5	2009	5	...	6	C4U-288	30	5	3	2	2	1	1	1.00
69016	1	3	9	1975	46	1	3	10	2011	2	...	9	S4-4672	1	8	8	2	3	1	1	1.00
1311191	1	31	12	2002	18	1	8	32	2012	8	...	9	D1H-465	1	2	4	4	1	0	0	1.00
1282689	1	6	11	1979	42	1	3	21	2015	2	...	9	2500-BC	1	2	4	5	1	1	1	1.00

479676 rows × 30 columns

```
# Finaliza entrenamiento
```

Figura 19
Predicciones del Modelo
Fuente: Pycarest – Python

Observamos que nuestro modelo nos revela las predicciones en nuestro nuevo dataset, para entrenamiento y test, en este nuevo conjunto de datos vemos dos nuevas columnas como son “prediciton_label” y “prediction_score”, con datos boleanos para acertar las prediciones entre 0 y 1.

```
Transformation Pipeline and Model Successfully Saved
(Pipeline(memory=Memory(location=None),
          steps=[('date_feature_extractor',
                  TransformerWrapper(exclude=None, include=['Fecha_Emision_Cat'],
                                     transformer=ExtractDateTimeFeatures(features=['day',
                                                                                   'month',
                                                                                   'year']))),
                 ('numerical_imputer',
                  TransformerWrapper(exclude=None,
                                     include=['Sexo', 'dia_nac', 'mes_nac',
                                              'anno_nac', 'Edad', 'Tipo_Persona',
                                              'ID_MARCA', 'Modelo', 'ann...
                                                      min_samples_leaf=20,
                                                      return_df=True,
                                                      smoothing=10,
                                                      verbose=0))),
                 ('actual_estimator',
                  DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None,
                                         criterion='gini', max_depth=None,
                                         max_features=None, max_leaf_nodes=None,
                                         min_impurity_decrease=0.0,
                                         min_samples_leaf=1, min_samples_split=2,
                                         min_weight_fraction_leaf=0.0,
                                         random_state=123, splitter='best'))],
          verbose=False),
 'modelo_propension_v1.pkl')
```

Figura 20
Modelo de predicción
Fuente: Pycarest – Python

Podemos observar en la figura 19 que nuestro modelo de predicción utilizando el algoritmo árbol de decisión o **DecisionTreeClassifier** realiza la caracterización de los campos tipo date o fechas y los campos de tipos categóricos, asignándole valores a cada nodo en donde encuentra 0 ó 1 según la valorización de la columna antes mencionado.

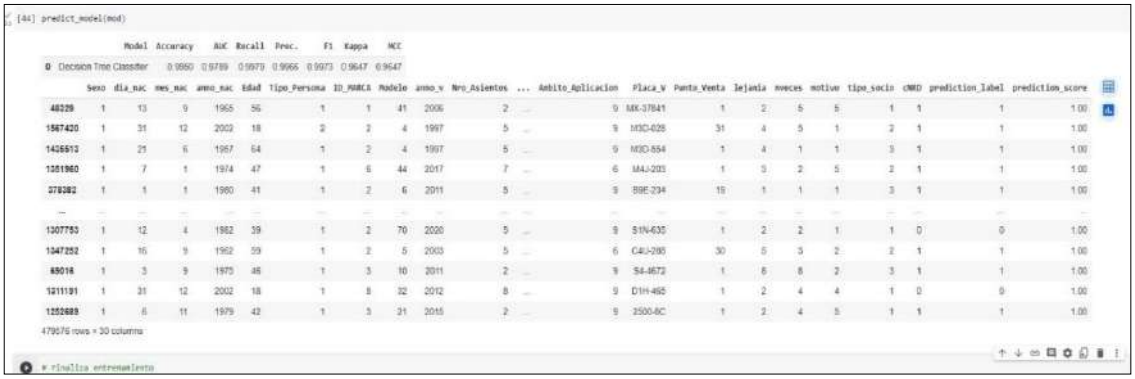
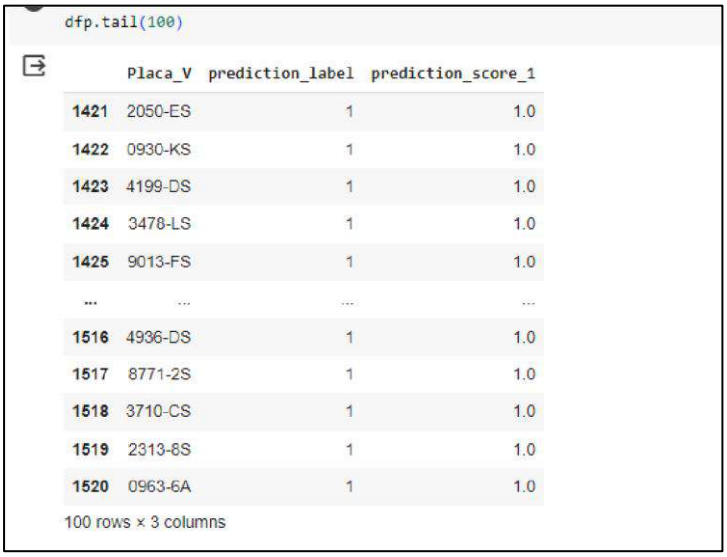


Figura 21
DataSet de Predicciones
Fuente: Pycarest – Python

Observamos en la figura 21, que se añadió a nuestro data set la columna “prediction_score”, que es la columna que nos indica quienes realizaran una posible afiliación, esta columna esta indexada por la placa del vehículo, que está relacionado a un asociado.

Realizamos la carga de nuestra data para realizar las predicciones de nuestro modelo entrenado, en este caso de los meses donde se evaluará el Modelo de machine learnign y su posterior evaluación detallado en los pasos siguientes.



```
dfp.tail(100)
```

	Placa_V	prediction_label	prediction_score_1
1421	2050-ES	1	1.0
1422	0930-KS	1	1.0
1423	4199-DS	1	1.0
1424	3478-LS	1	1.0
1425	9013-FS	1	1.0
...	...	...	...
1516	4936-DS	1	1.0
1517	8771-2S	1	1.0
1518	3710-CS	1	1.0
1519	2313-8S	1	1.0
1520	0963-6A	1	1.0

100 rows × 3 columns

***Figura* 22**
Muestra Resultados
Fuente: Pycarest – Python

Observamos que nuestros resultados en las etiquetas de las columnas “prediction_label” y “prediction_score_1” acertando en el pronóstico de afiliación o compra.

5. implementación y monitoreo del modelo diseñado para el logro del objetivo específico planteado.

Para este paso después de haber encontrado nuestro modelo entrenado y realizar las pruebas respectivas evaluamos el resultado obtenido, en este caso la data que nos resulta como posible comprador, se carga en el sistema de ventas en la opción de llamadas, siendo envidaos a los encargados de realizar la llamada el proceso final de la implementación.

6. Comparación del resultado obtenido con modelo de machine learning y resultado obtenido sin modelo.

Para la investigación entonces realizamos la toma de datos antes de la implementación según conveniencia del investigador por 30 días, y después de tabular y anotar los

resultados con su respetiva interpretación son separados para su posterior comparación.

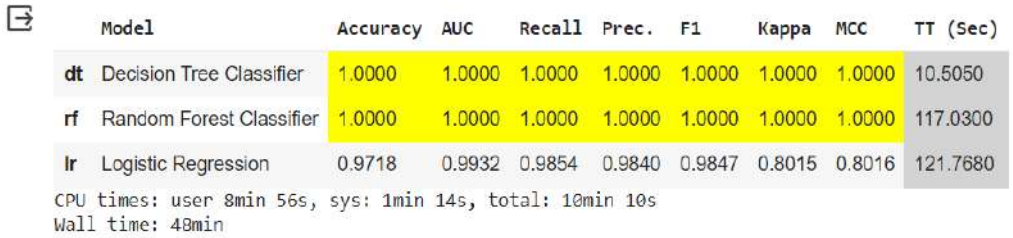
Después pasamos a la evaluación con la implementación del modelo por otros 30 días, realizando la captura de los datos y resultado obtenidos para su comparación y prueba de hipótesis respectiva.

3.3.2 Objetivo específico 2: Pronosticar la Estimación de las coberturas con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin

Se desarrollaron las siguientes actividades:

1. Recopilación de datos de los reportes de pagos realizado a los beneficiarios, coberturas, personas, centros médicos, reportes a la SBS, estados financieros, publicaciones y archivos, tanto digitales y físicos, que se tienen desde que la AFOCAT inicio sus operaciones, desde el 2008 hasta el 2022, que es donde se efectuó la investigación.
2. Limpieza y preparación de datos usando metodología KDD, que la poseer data de distintas fuentes, como archivos xls, txt, vsc, pdf, img y back-up de motores de base de datos como Oracle, Sql Server 2008 R2, MySql Server y MS Acces.
3. Selección del modelo de Machine Learning que según los datos y objetivo que queremos lograr determinamos el algoritmo de aprendizaje automático supervisado como modelo que ejecutamos.

Para este caso realizamos la evaluación de los gastos de siniestros incurridos durante los últimos periodos, y cargarlos a nuestro cuaderno de trabajo en Google Colaboratory, obteniendo el siguiente resultado.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	10.5050
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	117.0300
lr	Logistic Regression	0.9718	0.9932	0.9854	0.9840	0.9847	0.8015	0.8016	121.7680

```
CPU times: user 8min 56s, sys: 1min 14s, total: 10min 10s
Wall time: 48min
```

Figura 23
Modelo de machine learning - siniestros
Fuente: Pycarest – Python

Observamos que así como para el caso del objetivo anterior el mejor algoritmo se aplica con Decision Tree Classifier o con Random Forest Classifier, en este caso nos decantaremos por el más conocido y ya trabajado en los puntos anteriores, **Decisión Tree Classifier**

4. Entrenamiento del modelo y Validación del modelo realizado para la investigación, esto lo realizamos en el entorno de desarrollo Anaconda en cuadernos de trabajo de jupyter notebook con el lenguaje de programación Python en su versión 3.11.4, con los datos del objetivo anterior.

5. implementación y monitoreo del modelo diseñado para el logro del objetivo específico planteado.

Para este paso después de haber encontrado nuestro modelo entrenado y realizar las pruebas respectivas evaluamos el resultado obtenido, en este caso la data que nos resulta como posible estimación de gastos de siniestros, se carga en el sistema de siniestro para que se realice la comparación en donde se obtengan aciertos y pasara procesarlos como gastos incurridos por el siniestro, siendo envidaos a los encargados de realizar el proceso final de la implementación.

6. Comparación del resultado obtenido con modelo de machine learning y resultado obtenido sin modelo.

Para la investigación entonces realizamos la toma de datos antes de la implementación según conveniencia del investigador por 30 días, y después de tabular y anotar los resultados con su respetiva interpretación son separados para su posterior comparación.

Después pasamos a la evaluación con la implementación del modelo por otros 30 días, realizando la captura de los datos y resultado obtenidos para su comparación y prueba de hipótesis respectiva.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1 "Resultado específico 1: Gestionar el incremento de la cantidad de las afiliaciones con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin".

La investigación se realizó en 2 etapas, una primera etapa tomando los datos de 30 días calendarios en referencia las ventas realizadas sin la implantación del modelo de machine learning, donde se tomaron medidas a los indicadores de cada objetivo planteado , y la segunda etapa con datos de ventas de 30 días calendarios con la implantación del modelo de machine learning, volviendo a tomar los datos de los indicadores de cada objetivo específico, los analizamos y presentamos los siguientes resultados

Tabla 4

Fechas de evaluación

Etapa	Fecha de inicio	Fecha de fin
Sin – ML	01-06-2022	30-06-2022
Con - ML	01-07-2022	30-07-2022

Fuente: Elaboración propia del autor

Tomamos las ventas de 30 días y realizamos un análisis exhaustivo en sus distintos indicadores, ventas, numero personas, edad personas, sexo personas, dirección, tipo vehículo, tipo seguro, rango de edades, dirección, fecha de afiliación, etc.

Tabla 5

Medidas del indicador 1. Afiliaciones

Etapa	N	Mínimo	Máximo	Media	Desviación Estándar
Sin – ML	30	8	98	51.7	23.66
Con - ML	30	16	109	61	24.89
N Valido	30				

Fuente: Elaboración Propia del autor

En la tabla 3 se muestran las medidas estadísticas del indicador 1 de afiliaciones, donde podemos observar que los datos obtenidos reflejan que la media obtenida aumenta de 51.7 afiliaciones (Sin ML) a 61 afiliaciones (Con ML) evidenciando que con el Modelo de Machine learning se aumenta las afiliaciones en 10 afiliaciones (21 %).

En los extremos de afiliaciones se presentó aumento de la evaluación Sin-ML de 8 a 98 a la evaluación Con-ML de 16 a 109, aumentando de 90 a 93 afiliaciones y en referencia a la desviación estándar en la evaluación Sin-ML ± 23.66 y en la evaluación Con-ML ± 24.89, validando el incremento mencionado.

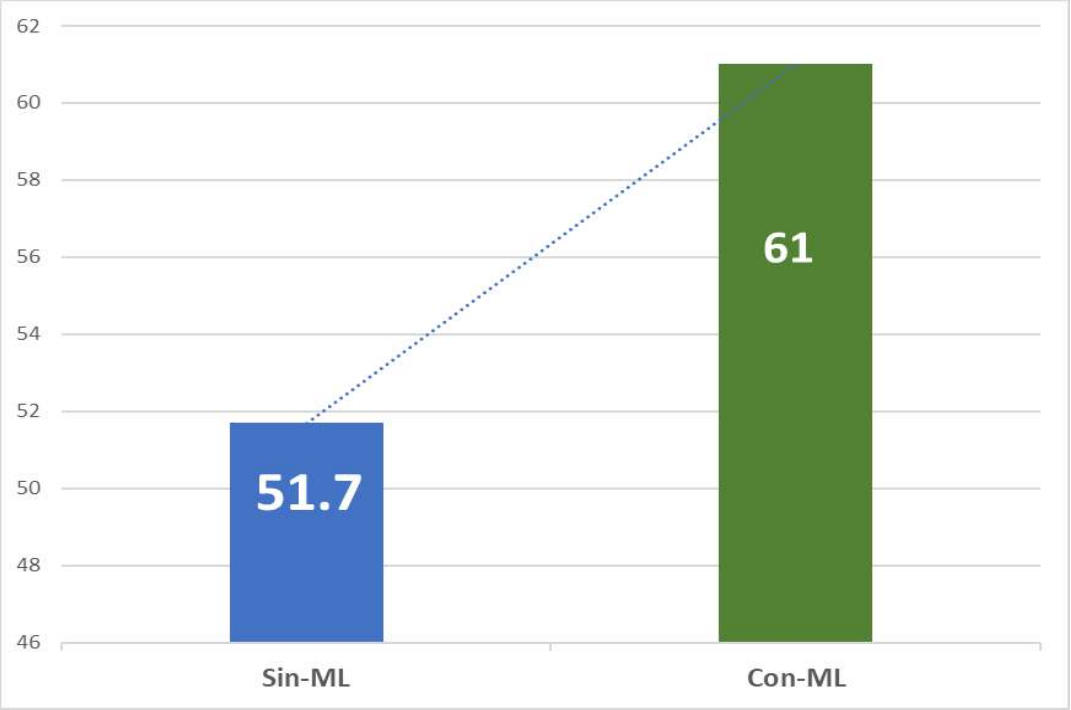


***Figura* 24**
Medidas del indicador afiliaciones
Fuente: Elaboración Propia

En la figura anterior podemos observar el incremento de afiliaciones utilizando un modelo de machine learning en un 21 % en Afocat San Martin en el 2022.

***Tabla* 6**
Pruebas de Normalidad del Indicador I

Pruebas de normalidad

	"Kolmogorov-Smirnov[a]"			"Shapiro-Wilk"		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
SinML	.171	30	.025	.934	30	.06396
ConML	.177	30	.017	.930	30	.12220

Fuente: Elaboración Propia

La prueba de normalidad mediante Shapiro-Wilk reveló que el valor de probabilidad (p) en ambos casos superaba el umbral de significación seleccionado (0,05). En consecuencia, se deduce que los datos tienen una distribución normal. Por este motivo, utilizamos una prueba paramétrica, en especial la prueba t de Student, para nuestra investigación.

Prueba de Hipótesis

***Tabla* 7**

Hipótesis para el indicador – Gestionar el incremento de la cantidad de las afiliaciones con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin.

Indicador 1	**“Gestionar el incremento de la cantidad de las afiliaciones con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin”.**
Ho: La implementación de un modelo de ML NO incrementa las afiliaciones en AFOCATSM Hi: La implementación de un modelo de ML incrementa las afiliaciones en AFOCATSM	
Ho: Sin-ML - Con-ML =< 0 Hi: Sin-ML - Con-ML > 0	
donde: Sin-ML: Afiliaciones ocurridas sin la implementación de un modelo de ML, Con-ML: Afiliaciones ocurridas con la implantación de un modelo de ML	

Fuente: Elaboración Propia

Para aplicar la prueba T de Student y evaluar la hipótesis planteada para este indicador, se utilizó un nivel de confianza del 95%. Se seleccionó un valor Z de 1.96, correspondiente a un margen de error del 5%.

***Tabla* 8**

Correlaciones de muestras emparejadas para el indicador I

“Correlaciones de muestras emparejadas”

		N	Correlación	Significación	
				P de un factor	P de dos factores
Par 1	SinML & ConML	30	.173	.180	.359

Fuente: Elaboración Propia

Cómo se evidencio anteriormente durante la investigación se obtuvieron datos paramétricos normales, siendo necesario realizar la mencionada prueba T-student.

***Tabla* 9**

Prueba de muestras emparejadas para el indicador I

		Diferencias emparejadas								
		Media	Desv. estándar	Media de error estándar	95% de intervalo de confianza de la diferencia		t	gl	Significación	
					Inferior	Superior			P de un factor	P de dos factores
Par 1	SinML - ConML	-10.467	28.846	5.267	-21.238	.305	-1.987	29	.028	.056

Fuente: Elaboración Propia

En las observaciones realizadas sobre los datos obtenidos tenemos que valor de t = -1.987< 1,96 a si mismo el p (Sig < 0.05), ante estos resultados obtenidos en el indicador podemos decir la hipótesis nula es rechazada, confirmando que La implementación de un modelo de ML se logra gestionar el incremento de las afiliaciones en Afocat San Martin en 2022.

4.2 "Resultado específico 2: Pronosticar la Estimación de gastos ocurridos en las coberturas con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin"

Después de revisar la data de siniestros reportados, de los cuales se determinan según el tipo de cobertura requerida, entre los años 2014 y 2021, nos resulta la siguiente tabla.

***Tabla* 10**
Promedio de costo total de siniestros por años

Años	Promedio de Costo Total de Siniestro
2014	S/ 1,005.50
2015	S/ 910.80
2016	S/ 2,012.47
2017	S/ 4,134.64
2018	S/ 6,052.72
2019	S/ 7,511.61
2020	S/ 9,279.92
2021	S/ 10,352.27
Total	S/ 41,259.93

Fuente: Elaboración Propia

Promedio de pronósticos de costo total de siniestros

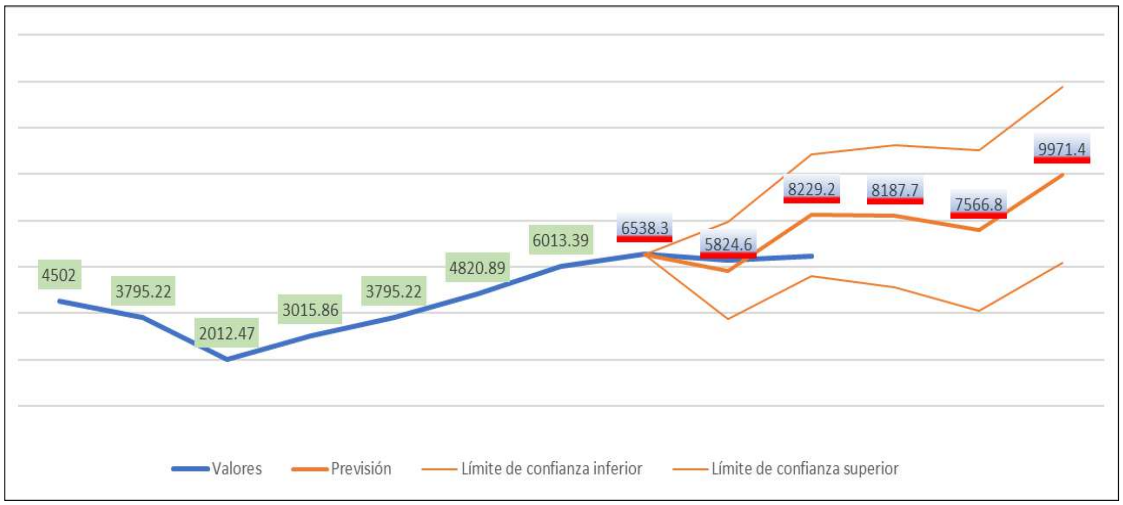


Figura 25
Pronóstico de costos
Fuente: Elaboración Propia

Tabla 11
Fechas de evaluación de estimación de costo de siniestros

Etapa	Fecha de inicio	Fecha de fin
Sin – ML	01-01-2021	31-12-2022
Con - ML	01-01-2021	31-12-2022

Fuente: Elaboración propia del autor

Tomamos los datos de nuestro pronostico, junto con los datos obtenidos para validar nuestra hipótesis.

Tabla 12
Medidas del indicador 2. Predecir gastos de siniestros

Etapa	N	Mínimo	Máximo	Media	Desviación Estándar
Sin – ML	24	2507.1	10845.25	6584.81	2226.45
Con - ML	24	6585.6	31277.5	18959.79	1232.29
N Valido	24				

Fuente: Elaboración Propia del autor

En la tabla N° 12, se muestran las medidas estadísticas del indicador 2 de gastos de siniestros, donde podemos observar que los datos obtenidos reflejan que la media obtenida aumenta de 6584.81(Sin ML) a 8959.79 (Con ML) evidenciando que con el Modelo de Machine learning se aumenta los aciertos en pronósticos en 96%.

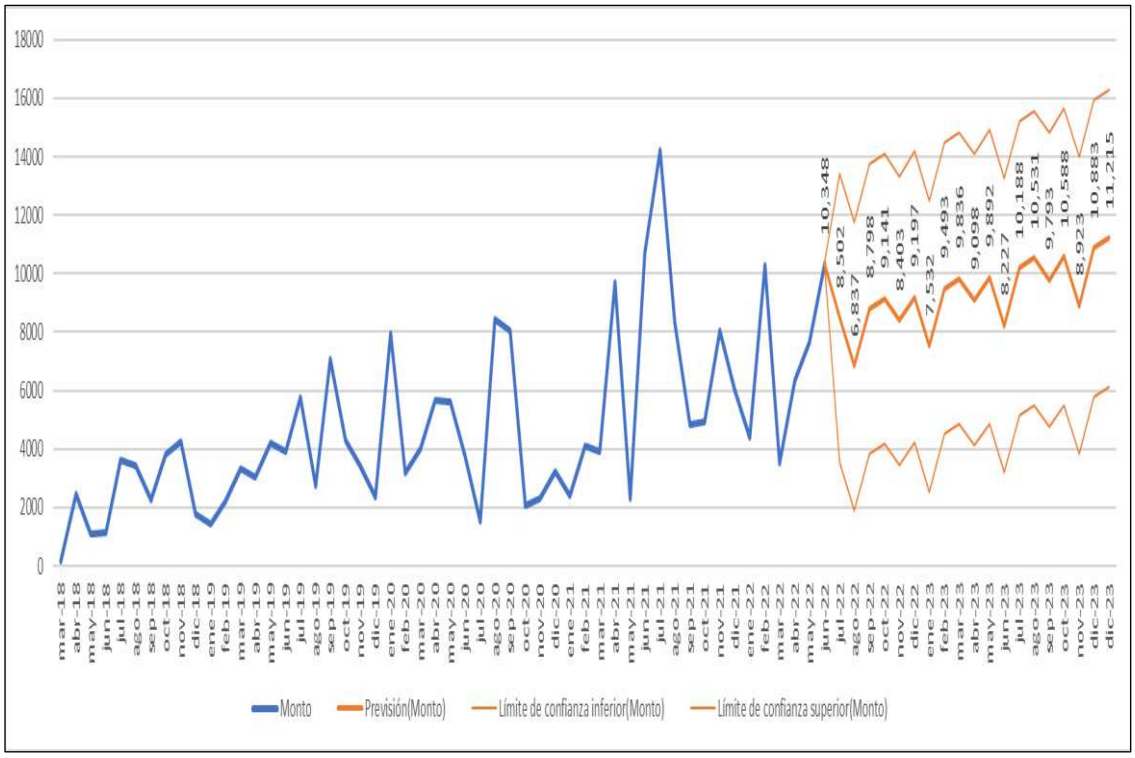


Figura 26
Pronósticos
Fuente: Elaboración Propia

En el grafico anterior podemos observar que nuestra modelo de machine logra acertar en todas las predicciones en 96 % de certeza, en comparación con los resultados obtenidos para los siguientes periodos para el año 2022.

Tabla 13
Pruebas de normalidad del indicador 2

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**	**Estadístico**	**gl**	**Sig.**
Sin_ML	.116	24	.200*	.928	24	.086
Con_ML	.057	24	.200*	.989	24	.995

Fuente: Elaboración Propia

Los resultados de la prueba de normalidad utilizando Shapiro-Wilk mostraron que el valor de probabilidad (p) en ambos casos fue superior a nuestro nivel de significancia seleccionado (0,05). Esto indica que los datos siguen una distribución normal. Por lo tanto, optamos por utilizar una prueba paramétrica, específicamente la prueba t de Student, para verificar nuestra hipótesis en este estudio.

Prueba de Hipótesis

Tabla 14

Hipótesis para el indicador – Pronosticar la Estimación de gastos ocurridos en las coberturas con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin

Indicador 2	**“Pronosticar la Estimación de gastos ocurridos en las coberturas con la implementación de un modelo basado en Machine Learning en AFOCAT San Martin”**
Ho: La implementación de un modelo de ML NO pronostica la estimación de gastos ocurridos en las coberturas de siniestros en AFOCAT San Martin Hi: La implementación de un modelo de ML NO pronostica la estimación de gastos ocurridos en las coberturas de siniestros en AFOCAT San Martin	
Ho: Sin-ML – Con-ML =< 0 Hi: Sin-ML – Con-ML > 0	
donde: Sin-ML: Estimación de gastos ocurridos por las coberturas de siniestros sin la implementación de un modelo de ML Con-ML: pronóstico de las Estimación de gastos ocurridos por las coberturas de siniestros sin la implementación de un modelo de ML	

Fuente: Elaboración Propia

Para realizar la prueba t de Student y contrastar la hipótesis planteada para este indicador, se tomó en cuenta un nivel de confianza del 90%. Se estableció un valor crítico de 1.645 para el intervalo de confianza del 90%, considerando un margen de error del 5%.

Tabla 15

Correlaciones de muestras emparejadas para el indicador 2

		N	Correlación	Significación	
				P de un factor	P de dos factores
Par 1	Sin_ML & Con_ML	24	-.112	.301	.601

Fuente: Elaboración Propia

Cómo se evidencio anteriormente durante la investigación se obtuvieron datos paramétricos normales, siendo necesario realizar la mencionada prueba T-student.

Tabla 16

Prueba de muestras emparejadas para indicador 2

		Diferencias emparejadas							Significación	
		Media	Desv. estándar	Media de error estándar	95% de intervalo de confianza de la diferencia		t	gl	P de un factor	P de dos factores
					Inferior	Superior				
Par 1	Sin_ML - Con_ML	-2374.95	2663.08	543.59	-3499.46	-1250.42	-4.369	23	<.001	<.001

Fuente: Elaboración Propia

En las observaciones realizadas sobre los datos obtenidos tenemos que valor de t = -1.987< 1,96 a si mismo e el p (Sig < 0.05), ante estos resultados obtenidos en el indicador podemos decir la hipótesis nula es rechazada, confirmando pronóstico de las Estimación de gastos ocurridos por las coberturas de siniestros sin la implementación de un modelo de ML

4.3 Discusión

Los resultados obtenidos en la investigación presente, demuestra y evidencia que la utilización de machine learning en procesos de ventas y siniestros se logra el Gestionar el incremento y acertar en la predicción, con resultados significativos dándose la aprobación de las hipótesis especificas planteadas.

El resultado clave proporcionado es el valor de t, que se calculó como -1.987, y se compara con un valor crítico de 1.96. Usualmente, en las pruebas de hipótesis, este valor crítico se asocia con un nivel de confianza del 95%.

Dado que el valor de t (-1.987) es menor que el valor crítico (-1.96), esto sugiere que los resultados son estadísticamente significativos. Además, se menciona que el valor de p (probabilidad) asociado con este valor de t es menor que 0.05 (p < 0.05). En términos generales, un valor de p menor que 0.05 se interpreta como evidencia suficiente para rechazar la hipótesis nula.

Lo discutido en el párrafo anterior, se complementa con lo concluido en la investigación publicada por Aceituno (2019), donde se logra determinar el asertividad para la captación de microcréditos, logrando reducir el riesgo de los otorgamientos de estos.

Por lo tanto, basándonos en estos resultados estadísticos, podemos concluir que la hipótesis nula es rechazada. Esto implica que existe evidencia estadística para

respaldar la afirmación de que la implementación de un modelo de ML incrementa las afiliaciones en Afocat San Martin. En otras palabras, se demuestra que el uso de este modelo tiene un efecto positivo y significativo en el aumento de las afiliaciones en la entidad.

En nuestra investigación obtuvimos datos que reflejan que la media obtenida aumenta de 51.7 afiliaciones (Sin ML) a 61 afiliaciones (Con ML) evidenciando que con el Modelo de Machine learning se aumenta las afiliaciones en 10 afiliaciones (21 %).

En los extremos de afiliaciones se presentó aumento de la evaluación Sin-ML de 8 a 98 a la evaluación Con-ML de 16 a 109, aumentando de 90 a 93 afiliaciones y en referencia a la desviación estándar en la evaluación Sin-ML ± 23.66 y en la evaluación Con-ML ± 24.89, validando el incremento mencionado.

Estos resultados se complementan con la investigación realizada por Montilla (2022), donde concluye que la utilización de aprendizaje automatizado predice el rendimiento académico con indicadores del 95 % de certeza dejando en evidencia el uso de las mencionadas herramientas.

Es importante tener en cuenta que estos resultados son específicos para el contexto y los datos del estudio en cuestión. No implican necesariamente una causalidad directa, sino más bien una asociación estadística significativa. Además, es fundamental considerar las limitaciones del estudio y la calidad de los datos utilizados para llegar a estas conclusiones.

En las observaciones realizadas sobre los datos obtenidos en nuestro segundo indicador, se ha calculado un valor de t igual a -1.987. Este valor se compara con un valor crítico de 1.96, que generalmente se asocia con un nivel de confianza del 95% en una prueba de hipótesis. La hipótesis nula en este caso generalmente sería que no hay diferencia significativa entre los resultados observados y los resultados esperados bajo ciertas condiciones. Sin embargo, dado que el valor de t (-1.987) es menor que el valor crítico (-1.96), esto sugiere que los resultados son estadísticamente significativos, lo que implica que existe una diferencia significativa entre los resultados observados y los esperados bajo la hipótesis nula.

Además, se menciona que el valor de p (probabilidad) asociado con este valor de t es menor que 0.05 ($p < 0.05$). En la mayoría de los casos, un valor de p menor que 0.05 se interpreta como evidencia suficiente para rechazar la hipótesis nula. Esto significa que

hay suficiente evidencia estadística para rechazar la idea de que no hay diferencia significativa.

En consecuencia, se concluye que la hipótesis nula es rechazada, lo que respalda la afirmación de que la implementación de un modelo de Machine Learning (ML) tiene un impacto significativo en la estimación de los gastos ocurridos por las coberturas de siniestros. En otras palabras, parece que el uso de un modelo de ML mejora la precisión de la estimación de gastos en comparación con un enfoque no basado en ML.

CONCLUSIONES

1. Como principal conclusión debemos indicar que con la implementación del modelo de machine learning se logró llegar a los objetivos planteados de la investigación, aumentar las afiliaciones y predecir los gastos por siniestros ocurridos, en Afocat San Martin en 2022, logrando evidenciar y encontrar resultados significativos relevantes para los objetivos específicos.

2. Se evidencia que con la utilización de un modelo de machine learning se logró gestionar el incremento de las afiliaciones reflejando que la media obtenida aumenta de 51.7 afiliaciones (Sin ML) a 61 afiliaciones (Con ML) evidenciando que con el Modelo de Machine learning se aumenta las afiliaciones en 10 afiliaciones (21 %) en Afocat San Martin en 2022.

3. Se concluye así también que con la utilización del modelo de machine learning logramos predecir los costos de siniestros en Afocat San Martin, donde obtuvimos que la media obtenida aumenta de 6584.81(Sin ML) a 8959.79 (Con ML) evidenciando que con el Modelo de Machine learning se aumenta las afiliaciones en 35 %. logrando acertar en todas las predicciones en 96 % de certeza, en comparación con los resultados obtenidos para los siguientes periodos para el año 2022 en la gestión de siniestros en Afocat San Martin en 2022.

4. otra conclusión que logramos destacar es que la utilización del modelo de machine learning ayudamos a lograr los objetivos de gestión trazados por las metas, misión y visión de la empresa Afocat San Martin en 2022.

RECOMENDACIONES

1. Se recomienda el gestionar en la empresa Afocat San Martin en 2022, un adecuado almacenamiento de los datos, tanto físicos y digitales, salvaguardando mediante políticas de administración de seguridad de la información, logrando que los datos mantengan su integridad, información tanto física y lógica.

2. Como recomendación para los gerentes de informática de Afocat San Martin, debemos garantiza que el modelo siga funcionando, alimentando los datos necesarios y válidos, para que el modelo siga aprendiendo con más variables y más data.

3. se recomienda en la gestión de personal que pueda desarrollar más alternativas de implementación de modelos de aprendizaje automático, donde se pueda desenvolver sin ninguna restricción, económica, recursos y generación de conocimiento, que es muy útil para la institución y sirve para conocer a través de los datos a sus afiliados y los gastos en los que se incurren para cada uno de los siniestrados.

4. se recomienda que con cada descubrimiento ocurrido al momento del análisis de datos deban ser utilizados para creación de productos y políticas que ayuden a los afiliados y siniestrados, como son los tipos de clientes, el rango de edades, tipos de gastos, tipos de coberturas, montos y productos farmacéuticos utilizados en la gestión de siniestros de Afocat San Martin.

5. Para futuras investigaciones se recomienda unir datos de otras empresas aseguradoras en referencia a sus estados financieros y afiliaciones, en tanto no infrinja políticas de privacidad, y generar un modelo nacional para mejorar las políticas de captación de clientes y estimación de gastos de siniestros en cada una de las empresas aseguradoras del país.

REFERENCIAS BIBLIOGRÁFICAS

Loudly soundtracks. (2024). Loudly.com. Recuperado el 3 de julio de 2024, de https://www.loudly.com/music/ai-music-generator

Arévalo, A. A. (2022). *Modelo de Gestión Publicitario Basado en Inteligencia Artificial para la Escuela de Posgrado de la Universidad Nacional de San Martín*. https://repositorio.unsm.edu.pe/. https://repositorio.unsm.edu.pe/bitstream/11458/4488/1/TESIS%20MAESTRIA%20ALBERTO%20ALVA%20AR%c3%89VALO.pdf

Aceituno Rojo, M. R. (2029). *Modelo predictivo de análisis de riesgo crediticio usando Machine Learning en una entidad del sector microfinanciero*. alicia.concytec.gob.pe. https://alicia.concytec.gob.pe/vufind/Record/RNAP_d37b636d16b3ae6c1d6e9dc3d2011839

Caselli Gismondi, H. E. (2021). *MODELO PREDICTIVO BASADO EN MACHINE LEARNING COMO SOPORTE PARA EL SEGUIMIENTO ACADÉMICO DEL ESTUDIANTE UNIVERSITARIO*. Repositorio UNS. https://repositorio.uns.edu.pe/bitstream/handle/20.500.14278/3804/52337.pdf?sequence=5&isAllowed=y

Documentación de Google Cloud. (2023). Google Cloud. Recuperado el 5 de julio de 2024, de https://cloud.google.com/docs?hl=es-419

La Fuente Carmona, D. (2022). *Diseño de soluciones avanzadas basadas en técnicas de machine learning para la toma de decisiones en gestión de activos*. https://idus.us.es/handle/11441/135570

Bernedo, J. F. M., Bustamante, F. J. S., & Yovera, R. S. V. (2021). *Identificación de Obras Urbanas para la Ciudad de Lima a Través del uso de Herramientas Basadas En Machine Learning*. https://www.proquest.com/openview/dd374dd1cc1de1fe2bed1d5d014b26fa/1?pq-origsite=gscholar&cbl=2026366&diss=y

Montilla Garcia, H. (2022). *"Algoritmos de aprendizaje automático supervisado en la predicción del rendimiento académico"*. renati.sunedu.gob.pe. https://renati.sunedu.gob.pe/handle/sunedu/3645348

Guzman Velez, D. M. (2022). *PROGRAMA ACADÉMICO DE MAESTRÍA EN INGENIERÍA DE SISTEMAS CON MENCIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN*. Universidad Cesar Vallejo. https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/102577/Guzman_VDM-SD.pdf?sequence=4&isAllowed=y

Géron, A. (2020). *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow: Conceptos, herramientas y técnicas para construir sistemas inteligentes*.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Raschka, S., Liu, Y. H., & Mirjalili, V. (2021). *Machine Learning Con Pytorch Y Scikit-Learn* (1a ed.). alphaeditorial marcombo.

Flach, P. (2012). *Machine Learning: The art and science of algorithms that make sense of data* (Vol. 396). Cambridge University Press.

Goodfellow, I. Kurakin, A., y Bengio, S. (2016). Aprendizaje automático adversario a escala. *Preimpresión de arXiv arXiv:1611.01236* .

Capellman, J. (2020). *Hands-On Machine Learning with ML.NET: Getting started with Microsoft ML.NET to implement popular machine learning algorithms in C#*. Packt Publishing.

Bagnato, J. I. (s/f). *Aprende Machine Learning en Español: Teoría + Práctica Python (Spanish Edition)*.

Bengio, Y., Courville, A., & Vincent, P. (2012). Representation learning: A review and new perspectives. En *arXiv [cs.LG]*. http://arxiv.org/abs/1206.5538

Ministerio de Transportes y Comunicaciones. (s/f). Gob.pe. Recuperado el 4 de julio de 2024, de https://www.gob.pe/institucion/mtc/normas-legales/5188179-004-2024-mtc

Microsoft. (2023). Azure Machine Learning. Microsoft.com. https://azure.microsoft.com/es-es/products/machine-learning

Villatoro, F. R. (2017, octubre 11). *Geoffrey Hinton, el padre del aprendizaje profundo (deep learning)*. La Ciencia de la Mula Francis.

https://francis.naukas.com/2017/10/11/geoffrey-hinton-el-padre-del-aprendizaje-profundo-deep-learning/

Arroyo, M. J. (2022, enero 27). *Informe Digital 2022 - Cada vez más usuarios (y más enganchados) de Internet y las redes*. Linkedin.com. https://www.linkedin.com/pulse/informe-digital-2022-cada-vez-m%C3%A1s-usuarios-y-de-las-jim%C3%A9nez-arroyo/

Thompson, H. H. (2022, marzo 7). *INTELIGENCIA ARTIFICIAL Y PERIODISMO*. Observatorio de Tecnologías. https://perio.unlp.edu.ar/sitios/observatoriodetecnologias/inteligencia-artificial-y-periodismo/

Kemp, S. (2024). *Digital 2024 April global statshot report*. DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2024-april-global-statshot

ANEXOS

1. Caracterización del modelo de Datos

Nombre Columna	Definición
Sexo	sexo referido por el cliente al momento de la compra 1= Masculino 2= Femenino
dia_nac	día del nacimiento del socio, afiliado o agremiado y sin gremio al momento de la afiliación. Valor de tipo Categórico, esta entre 1 y 31
mes_nac	mes del nacimiento del socio, afiliado o agremiado y sin gremio al momento de la afiliación. Valor de tipo Categórico, esta entre 1 y 12
anno_nac	año del nacimiento del socio, afiliado o agremiado y sin gremio al momento de la afiliación.
Edad	Edad del nacimiento del socio, afiliado o agremiado y sin gremio al momento de la afiliación.
Tipo_Persona	identificador para el tipo de afiliado si es persona natural o personal jurídica. 1 = Persona Natural 2 = Persona Jurídica
ID_MARCA:	Marca del vehículo al que se le asigna el certificado contra accidentes de tránsito CAT. 1 = HONDA 2 = TOYOTA 3 = YAMAHA 4 = BAJAJ 5 = NISSAN 6 = SUZUKI 7 = WANXIN 8 = LIFAN 9 = HYUNDAI 10 = DAEWOO 11 = C & C MOTORS 12 = CHEVROLET 13 = MITSUBISHI 14 = KIA 15 = VOLKSWAGEN 16 = MAZDA 17 = RTM 18 = CHANGAN 19 = CHERY 20 = GREAT WALL 21 = JAC 22 = DAIHATSU 23 = FORD

	23 = VOLVO 23 = HAFEI 23 = DODGE 23 = GEELY 23 = FOTON 23 = RENAULT 23 = JINBEI 23 = PEUGEOT 23 = FAW 23 = JEEP 23 = TIANJIN FAW 23 = DFSK 23 = HERO 23 = ZOTYE 23 = DONGFENG 23 = NOTE 23 = DATSUN 23 = CITROEN 23 = SUBARU 23 = ASIA 23 =LADA 23 = KAYAK 23 = SKY MOTO 24 = ZONGSHEN 25 = QINGQI 26 = MOTOKAR 27 = CROSS 28 = APACHE
Modelo	Modelo relacionado a la marca del Vehículo al que se le asegura con el CAT 1 = CG L125 2 = WX 125 A 3 = XR125L 4 = COROLLA 5 = PROBOX 6 = YARIS 7 = STORM 8 = WAVE 9 = CB 125 S 10 = YB125 11 = SPRINTER 12 = PULSAR 13 = WAGON

	14 = ZS 15 = HILUX 16 = AVANZA 17 = FZ16 18 = SENTRA 19 = AM 150 FARMER ESPECIAL 20 = APV GL 21 = XTZ 22 = AG100 23 = CARINA 24 = C & C 110 25 = APACHE 26 = CRYPTON 27 = SUPER TICO 28 = PASSION 29 = ROUTER 30 = TIIDA 31 = ALTO 32 = LF 33 = PATROL 34 = LIBERO 35 = STROM 36 = AD 37 = I10 38 = TORNADO 39 = FT 150 GY-2 40 = RTM 41 = GL 125 42 = WY 125-A 43 = QM 44 = ERTIGA 45 = NEW VAN 46 = AT200GY HUNTER 47 = GOL GP COMFORT 1.6 48 = ACCENT 49 = EON 50 = moto de 2 asientos 51 = XR190L 52 = CB 110 53 = NLP 125 54 = XR150L 55 = XR190L 56 = XR250TORNADO 57 = HIACE DX 58 = CG125

	59 = XTZ 125 60 = XTZ-150 61 = NEW SUPERVAN 62 = GRAND VAN TURISMO 63 = FZS FI 64 = PASSAT 65 = TERIOS LONG 66 = PICANTO 67 = elantra 68 = WISH 69 = N300 70 = ETIOS 71 = RIO 72 = TERIOS 73 = TERCEL 99 = varios 100= Sin Datos 101= varios
anno_v	Año de fabricación del vehículo que se le asegura con el CAT
Nro_Asientos	Número de asientos del vehículo que se le asegura con el CAT.
Categoria_Vehiculo	Categoría del Vehículo a los cuales se les oferta el CAT. 0 = L4 1 = L5 2 = M1 3 = M2 4 = M3 5 =L3
Uso_vehiculo	Es el uso que se le da al vehículo que adquiere el CAT 1 = SERV. PÚB. URBANO 2 = INTERURBANO 3 = SERV. TRANSP. PÚB. REGIONAL 4 = SERV. TRANSP. PRIVADO 5 = CARGA 6 = SERV. TRANSP. TURÍSTICO 7 = no hay 8 = ESCOLAR Y TURISMO 9 = PRIVADO 10 = SERVICIO URBANO
Clase_Vehicular	clasificación de vehicular con el cual se le da a conocer en el momento de la afiliación y adquisición. 1 = CAMION 2 = CAMIONETA

	3 = AUTOMOVIL 4 = COMBI 5 = MINIVAN 6 = STATION WAGON 7 = OMNIBUS 8 = AMBULANCIA 9 = CAMIONETA PICK UP 10 = MOTO LINEAL 11 = MULTIPROPOSITO 12 = TRIMOVIL
Tipo_Carroceria	Tipo de carrocería del vehículo que adquiere el CAT. MOTO LINEAL 8 TRIMOVIL 6 AUTOMOVIL 1 CARGA 12 TRIMOVIL 2 MINIVAN 3 MULTIPROPOSITO 4
Fecha_Emision_Cat	Fecha de Emisión del CAT en Formato dd/mm/aaaa
dia_emision	Dia en el que se emite el CAT y es Afiliado el Socio o persona, puede ser del 1 al 31
mes_emision	mes en el que se emite el CAT y es Afiliado el Socio o persona, puede ser del 1 al 12
anno_emision	Año en el que se le emite el CAT
Hora_Emision	hora de emisión del CAT, se considera en formato de 24h, entre las 0 y 23 horas.
min_emision	Minuto de emisión del CAT, se consideran valores entre 1 y 31
cNRD	el cliente renovó o compro su CAT, esta será nuestra variable que nos determinará si el cliente compra o no su CAT 1 = Renovación de Certificado contra accidentes de tránsito (CAT) 0 = Nueva compra de Certificado contra accidentes de tránsito (CAT)
Ambito_Aplicacion	Provincia donde circulara el vehículo que adquirió el CAT 1 = San Martin 2 = Moyobamba 3 = bellavista 4 = Rioja 5 = Lamas 6 = Picota 7 = El Dorado 8 = Huallaga 9 = Mariscal Caceres

	10 = Tocache	
Placa_V	Placa del vehículo que adquiere el CAT	
Tipo_Socio	Tipo de afiliado que adquiere el CAT 1 = Socio 2 = Agremiado 3 = Sin Gremio	
Punto_Venta	Lugar donde adquirió el CAT 1 = OFICINA PRINCIPAL 2 = OFICINA SUCURSAL 3 = CACATACHI 4 = CAMPANILLA 5 = CERPA MOTORS 1 5 = CERPA MOTORS 2 5 = CERPA MOTORS 3 6 = JUANJUI II 7 = JUANJUI IV 8 = LAMAS I 9 = LAMAS II 10 = MORALES 1 11 = MOYOBAMBA 12 = NUEVA CAJAMARCA I 13 = NUEVA CAJAMARCA III 14 = NUEVO PROGRESO 17 = BELLAVISTA II 18 = BELLAVISTA III 19 = PICOTA 20 = PONGO DE CAYNARACHI 21 = RIOJA I 22 = RIOJA II 23 = RIOJA III 24 = RIOJA IV 25 = SAUCE 26 = SERTEL 27 = SERTEL 2 28 = SERVICIOS DRJ 29 = SISA 30 = SORITOR 31 = SURCO MOTORS 1,2,3,4,5,6,7 32 = TOCACHE I 33 = TOCACHE II 34 = UCHIZA I 35 = AUTOS CAJAMARCA 36 = EMTRATUR MOYOBAMBA 37 = NSP SAPOSOA	
lejanía	valor que considera el afiliado para llegar a adquirir su CAT	

	1= muy cerca menos de 1 km 2= cerca 3= cerca-lejos 4= lejos 5= lejos-lejano 6= lejano 7= muy lejano 8= muy lejano + 10km	
nveces	veces que el afiliado o persona adquirió su CAT 1= compro 1 vez 2= compro 2 veces 3= compro 3 veces 4= compro 4 veces 5= compro 5 veces 6= compro 6 veces 7= compro 7 veces 8= compro 8 veces 9= compro 9 veces 10 = compro 10 veces	
motivo	motivo por el cual el afiliado o persona adquiere su CAT 1 = ofertas 2 = necesidad 3 = vehículo nuevo 4 = precio 5 = otros	

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín en 2022

por Even Ronald Pérez Díaz

Fecha de entrega: 10-dic-2024 12:57p.m. (UTC-0500)
Identificador de la entrega: 2547956486
Nombre del archivo: TESIS_Even_Ronald_P_rez_D_az_10.12.2024.docx (3.67M)
Total de palabras: 15940
Total de caracteres: 88905

Modelo basado en machine learning para gestionar afiliaciones y siniestros en AFOCAT San Martín en 2022

INFORME DE ORIGINALIDAD

20%	19%	3%	10%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.unsm.edu.pe Fuente de Internet	3%
2	tesis.unsm.edu.pe Fuente de Internet	3%
3	hdl.handle.net Fuente de Internet	2%
4	es.wikipedia.org Fuente de Internet	2%
5	repositorio.uladech.edu.pe Fuente de Internet	1%
6	repositorio.ucv.edu.pe Fuente de Internet	1%
7	azure.microsoft.com Fuente de Internet	1%
8	www.coursehero.com Fuente de Internet	<1%
9	dokumen.pub Fuente de Internet	<1%
10	Submitted to Universidad Carlos III de Madrid Trabajo del estudiante	<1%
11	wiki2.org Fuente de Internet	<1%