UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERIA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Optimización Bayesiana de modelos de machine learning para mejorar la predicción de clientes e-learning

Tesis para obtener el Título Profesional de Ingeniero de Sistemas

Autor:

Jorge Daniel Maquera Canales

Asesor:

Magister Nemias Saboya Rios

Lima, setiembre 2023

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Nemias Saboya Rios, docente de la Facultad de Ingeniería y Arquitectura Escuela Profesional de Ingeniería de Sistemas , de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"OPTIMIZACIÓN BAYESIANA DE MODELOS DE MACHINE LEARNING PARA MEJORAR LA PREDICCIÓN DE CLIENTES E-LEARNING"** del autor Jorge Daniel Maquera Canales tiene un índice de similitud de 13% verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 28 días del mes de Setiembre del año 2023

NEMIAS SABOYA RIOS

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa Unión, a los **06** día(s) del mes de **setiembre** del año 2023 siendo **las 10:00 horas**, se reunieron en modalidad virtual u online sincrónica, bajo la dirección del Señor Presidente del jurado: **Dra. Erika Inés Acuña Salinas**, el secretario: **Mg. Geraldine Verónica Alvizuri Llerena**, y los demás miembros**: PhD. Javier Linkolk Lopez Gonzales** y **el MSc. Fredy Abel Huanca Torres**, y el asesor, **Mg. Nemias Saboya Rios**, con el propósito de administrar el acto académico de sustentación de la tesis titulada: " Optimización bayesiana de modelos de aprendizaje automático para mejorar la predicción de clientes e-learning "

de el(los)/la(las) bachiller/es: a) **JORGE DANIEL MAQUERA CANALES**

…… b) **……**

conducente a la obtención del título profesional de **INGENIERO DE SISTEMAS**
(Nombre del Título profesional)

con mención en……

El Presidente inició el acto académico de sustentación invitando al (los)/a(la)(las) candidato(a)/s hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del jurado a efectuar las preguntas, y aclaraciones pertinentes, las cuales fueron absueltas por el(los)/la(las) candidato(a)/s. Luego, se produjo un receso para las deliberaciones y la emisión del dictamen del jurado.

Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidato (a): …… **JORGE DANIEL MAQUERA CANALES** ……

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	19	A	Con nominación de Excelente	Excelencia

Candidato (b): …… **……** ……

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	

(*) Ver parte posterior

Finalmente, el Presidente del jurado invitó al(los)/a(la)(las) candidato(a)/s a ponerse de pie, para recibir la evaluación final y concluir el acto académico de sustentación procediéndose a registrar las firmas respectivas.

Presidente
Dra. Erika Inés Acuña Salinas

Secretario
Mg. Geraldine Veronica Alvizuri Llerena

Mg. Nemias Saboya Rios

Miembro
PhD. Javier Linkolk Lopez Gonzales

Miembro
MSc. Fredy Abel Huanca Torres

Candidato/a (a)
Jorge Daniel Maquera Canales

Candidato/a (b)

Índice

Optimización Bayesiana de modelos de machine learning para mejorar la predicción de clientes e-learning

Bayesian optimization of machine learning models to improve e-learning customer prediction

Resumen

La puntuación de clientes potenciales desempeña un papel crucial en el marketing al evaluar el nivel de interés y compromiso de posibles clientes. Las técnicas de aprendizaje automático ofrecen un medio para automatizar los procesos de puntuación de clientes potenciales, permitiendo a los profesionales del marketing priorizar sus esfuerzos y asignar recursos de manera efectiva. Sin embargo, el rendimiento de los modelos de aprendizaje automático depende en gran medida de la configuración de sus hiperparámetros. Los métodos de optimización tradicionales pueden ser ineficientes y no lograr explorar eficazmente el espacio de hiperparámetros de alta dimensionalidad. En este estudio, proponemos un enfoque novedoso que utiliza la optimización bayesiana de hiperparámetros para mejorar la predicción de conversión de clientes en el ámbito del e-learning. Al aprovechar las estadísticas bayesianas y un modelo probabilístico, nuestro método explora eficientemente el espacio de hiperparámetros para identificar configuraciones óptimas que maximizan el rendimiento del modelo. Consideramos los algoritmos de aprendizaje automático Extreme Gradient Boosting, Support Vector Machine, Random Forest, Logistic Regression y Decision Tree, y comparamos los algoritmos optimizados con sus versiones base en términos de precisión, sensibilidad, puntuación F1 y el área bajo la curva característica de operación del receptor (AUC). Los resultados demostraron que los algoritmos optimizados superaron consistentemente a sus versiones base. Nuestra investigación destaca la importancia de la optimización de hiperparámetros para lograr un rendimiento óptimo del modelo de aprendizaje automático y proporciona información valiosa para los profesionales del marketing en la industria del e-learning. Al aprovechar los algoritmos optimizados, las organizaciones pueden tomar decisiones basadas en datos, maximizar las tasas de conversión y optimizar las estrategias de marketing.

Palabras clave: puntuación de clientes potenciales, e-learning, optimización bayesiana.

Abstract

Lead scoring plays a crucial role in marketing by evaluating the level of interest and commitment of potential customers. Machine learning techniques offer a means to automate lead scoring processes, enabling marketers to prioritize their efforts and allocate resources effectively. However, the performance of machine learning models heavily relies on the configuration of their hyperparameters. Traditional optimization methods can be inefficient and fail to navigate the high-dimensional hyperparameter space effectively. In this study, we propose a novel approach using Bayesian hyperparameter optimization to enhance customer conversion prediction in the e-learning domain. By leveraging Bayesian statistics and a probabilistic model, our method efficiently explores the hyperparameter space to identify optimal configurations that maximize model performance. We considered the machine learning algorithms Extreme Gradient Boosting, Support Vector Machine, Random Forest, Logistic Regression, and Decision Tree, and compared the optimized algorithms against their base versions in terms of accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC). The results demonstrated that the optimized algorithms consistently outperformed their base

versions. Our research highlights the importance of hyperparameter optimization in achieving optimal machine learning model performance and provides valuable insights for marketers in the e-learning industry. By leveraging the optimized algorithms, organizations can make data-driven decisions, maximize conversion rates, and optimize marketing strategies.

Keywords: lead scoring, e-learning, Bayesian optimization

Introducción

Lead scoring es una práctica importante en el marketing, ya que consiste en evaluar el nivel de interés y compromiso de los clientes potenciales[1]. Durante su desarrollo se analiza el comportamiento en el sitio web, la interacción por correo electrónico, y la actividad en las redes sociales de los interesados para asignar una puntuación basada en la probabilidad de conversión[2]. El análisis de los datos y la identificación de los interesados más prometedores puede ser automatizado con aprendizaje automático[3]. De este modo, los responsables de marketing pueden priorizar sus esfuerzos y asignar recursos de forma más eficaz. Por lo cual se optimizarían las campañas de marketing, que resultarían en un mayor compromiso y conversión de los clientes potenciales[4].

Por otra parte, la inteligencia artificial y el aprendizaje automático son herramientas eficaces para la optimización de los procesos internos de una organización[5]. La toma de decisiones impulsada por datos, basada en inteligencia artificial (AI) o aprendizaje automático (ML), ofrece una mejor adaptabilidad al mercado altamente fluido, acelerando y mejorando la calidad de las decisiones[6]. Estos beneficios surgen a partir de que el proceso de decisión utiliza algoritmos de aprendizaje automático que realizan la extracción de información de grandes cantidades de datos[7]. Esta información puede impulsar y soportar la toma de una decisión de marketing[8].

Entre los factores que influyen el desempeño de un modelo de aprendizaje automático, tenemos la calidad de los datos, pero también la configuración de sus parámetros[9,10]. La optimización de los parámetros se realiza para que los algoritmos logren un desempeño óptimo[11] y consiste en encontrar la mejor configuración de forma automática[12]. Para aplicar esta técnica, primero se debe definir los parámetros que se combinarán con el fin de encontrar la mejor configuración. Después, se prueba cada una de estas, y se selecciona la que mejor funciona, a esta técnica se le llama GridSearch y requiere un alto costo computacional[13]. Una alternativa similar pero más eficiente es RandomSearch que no prueba todas las combinaciones sino un número determinado de combinaciones aleatorias[14]. Otro enfoque novedoso que ha demostrado un rendimiento superior a los métodos anteriores es la optimización bayesiana[15,16] que puede optimizar funciones costosas con gran precisión en poco tiempo[17,18].

Entretanto, una inadecuada optimización de hiperparámetros puede impactar de forma significativas al desarrollo de modelos de aprendizaje automático[9,19,20]. Los métodos tradicionales de optimización de hiperparámetros pueden ser altamente ineficientes y consumir mucho tiempo[12,21]. Además, comúnmente fallan en encontrar la configuración óptima de debido a su incapacidad de navegar el espacio de hiperparámetros de alta dimensión[6]. Otro factor que podría perjudicar el entrenamiento de algoritmos de ML es el empleo de no expertos, ya que se requiere un amplio conocimiento para ajustar con éxito los modelos[7]. Esto resulta en una pérdida de tiempo y recursos computacionales, así como una configuración inadecuada, que podría dar lugar a un bajo rendimiento del modelo[22]. Por otro lado, utilizar métodos más eficientes para encontrar la mejor configuración de hiperparámetros aportaría a alcanzar el mejor rendimiento de los modelos.

Optimizar los hiperparámetros es fundamental para lograr un rendimiento óptimo de modelos de aprendizaje automático. Sin embargo, el ajuste manual es a menudo poco práctico debido al gran espacio de parámetros. En este artículo, proponemos un enfoque novedoso para el ajuste de hiperparámetros utilizando la optimización bayesiana para mejorar la perdición de la conversión de los clientes de e-learning. Este enfoque utiliza la estadística bayesiana y un modelo probabilístico para explorar eficientemente el espacio de hiperparámetros y encontrar los ajustes óptimos que logren el mejor rendimiento del algoritmo de aprendizaje automático[23].

Además, puede reducir la incertidumbre en el proceso de decisión de clasificación de clientes potenciales. Las contribuciones específicas se detallan a continuación:

- La optimización bayesiana de los hiperparámetros puede mejorar el rendimiento de los modelos de predicción de conversión de clientes de e-learning en comparación con el uso de los valores predeterminados de los hiperparámetros.
- La evaluación del algoritmo optimizado que mejor reduzca la incertidumbre en el proceso de decisión de clasificación de clientes potenciales es comparada a otros algoritmos similares.

Fundamento teórico

Enfoque bayesiano

La estadística bayesiana es un enfoque probabilístico que utiliza el teorema de Bayes para actualizar nuestras creencias sobre la probabilidad de un suceso a partir de nuevas pruebas[24]. El teorema de Bayes establece que la probabilidad de una hipótesis (H) dados los datos (D) es proporcional al producto de la probabilidad a priori de la hipótesis ($P(H)$) y la probabilidad de los datos dada la hipótesis ($P(D|H)$), dividido por la probabilidad de los datos ($P(D)$). En otras palabras, podemos decir que la inferencia bayesiana deriva la probabilidad posterior a partir de una combinación de una probabilidad previa y una función de verosimilitud derivada de un modelo estadístico para los datos observados. Matemáticamente, esto se puede escribir como:

$$P(H|D) = P(D|H) \times P(H) \,/\, P(D).$$

Optimización bayesiana

La optimización bayesiana es una estrategia de diseño secuencial que se utiliza para optimizar funciones costosas de caja negra[25]. Se basa en la construcción de un modelo de probabilidad para la función objetivo, el cual se actualiza a medida que se evalúa la función. A partir de este modelo, se utiliza una función de adquisición para seleccionar la próxima configuración de entrada que tenga la mayor probabilidad de mejorar la mejor configuración actual. De esta manera, se evita explorar valores de entrada ineficaces, y se evalúan únicamente los valores de entrada más prometedores para la función objetivo. Este proceso se repite hasta encontrar la solución óptima o hasta alcanzar el límite de evaluaciones. Por lo tanto, la optimización bayesiana representa un método poderoso y eficiente para la optimización de hiperparámetros de funciones de alto costo, como la optimización de hiperparámetros de un modelo de aprendizaje automático[26].

Tree-Structured Parzen Estimator (TPE)

Tree-Structured Parzen Estimator (TPE) es un algoritmo de optimización bayesiana que se utiliza para construir el modelo de probabilidad[27]. TPE utiliza una distribución de densidad para modelar la probabilidad de que una configuración de hiperparámetros sea la óptima. La distribución se divide en dos partes, una que modela la probabilidad de que una configuración sea mejor que la actual mejor configuración, y otra que modela la probabilidad de que una configuración sea peor[13]. El algoritmo utiliza esta distribución para guiar la búsqueda de manera más efectiva.

Técnicas de aprendizaje automático

Este estudio aplicará algoritmos de aprendizaje automático para predecir si los usuarios de una plataforma e-learning se convertirán en clientes. A continuación, se explican brevemente los cinco algoritmos de aprendizaje automático que serán utilizados:

- Árbol de decisión (DT): Un árbol de decisión es un modelo en forma de árbol en el que cada nodo representa una característica, cada rama representa una regla de decisión basada en esa característica y cada hoja representa una predicción[28].
- Bosque aleatorio (RF): Un bosque aleatorio es un conjunto de árboles de decisión, donde cada árbol se construye sobre un subconjunto aleatorio de las características, y la salida es la moda de las predicciones individuales del árbol[29].
- Regresión logística (LR): La regresión logística es un modelo lineal de clasificación binaria que utiliza la función logística para asignar la combinación lineal de características a una probabilidad[30].
- Máquina de vectores soporte (SVM): Una SVM es un modelo lineal o no lineal que encuentra un hiperplano que separa al máximo las clases, basándose en los vectores de soporte[31].
- Extreme Gradient Boosting (XGB): XGB es un algoritmo de refuerzo que combina múltiples árboles de decisión débiles para crear un modelo fuerte, utilizando técnicas de refuerzo de gradiente y regularización para mejorar el rendimiento[32].

Propuesta

La metodología propuesta en este artículo se presenta en la Figura 1 y se explican sus detalles en las siguientes subsecciones.

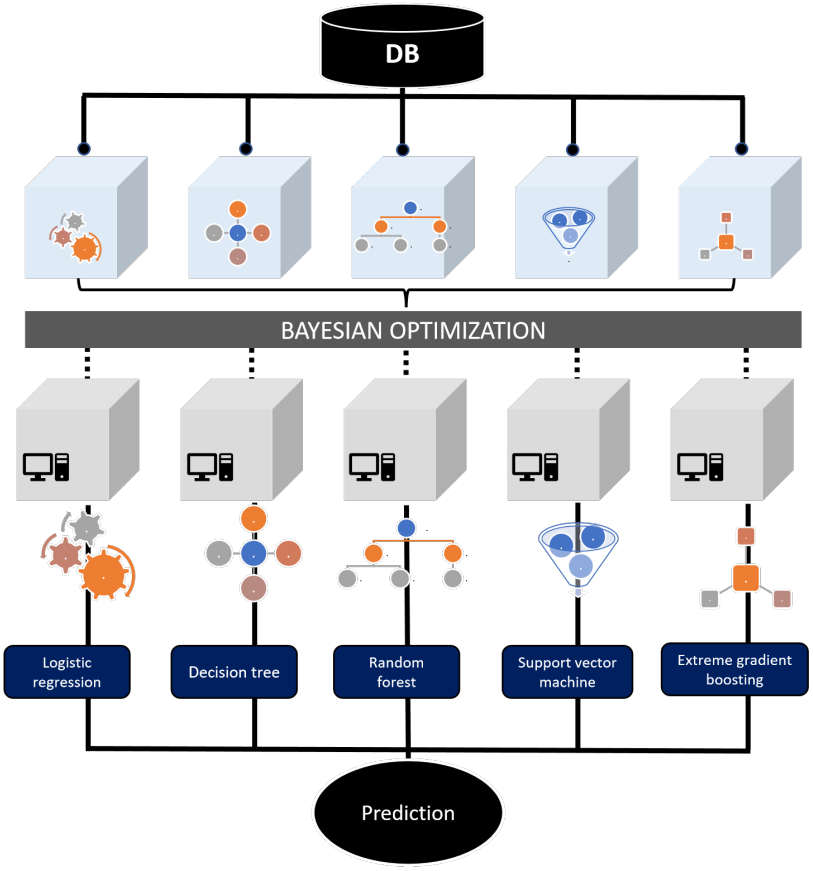

Figure 1. Estructura del procedimiento de optimización bayesiana propuesto en este artículo.

Comprensión de los datos

Descripción del conjunto de datos

Se utilizo el conjunto de datos libre "Lead Scoring X Online Education" que cuenta con 9240 registros y 37 columnas para la aplicación y para mostrar la utilidad de nuestra la propuesta[33]. Este contiene datos demográficos y sobre la conducta de los usuarios de una página web de e-learning. Los datos se prestan para realizar una clasificación binaria de los clientes que se dividen en convertidos y no convertidos. Las clases presentan un desequilibrio leve teniendo la clase minoritaria un 38% del conjunto de datos. En la Tabla 1 se describen las variables del conjunto de datos.

Estos datos también fueron empleados por otros autores que tenía el objetivo de demostrar los beneficios de usar algoritmos de aprendizaje automático en la automatización del proceso de puntuación para nuevos clientes potenciales con la implementación de modelos predictivos[34]. Aunque ambos estudios comparten el objetivo de mejorar la puntuación predictiva de clientes potenciales, difieren en cómo utilizan los datos. En nuestro estudio, utilizamos la optimización bayesiana para ajustar los hiperparámetros de los modelos de aprendizaje automático y así mejorar la precisión predictiva. Por otro lado, el otro estudio utiliza modelos predictivos para automatizar el proceso de puntuación para nuevos clientes potenciales, sin enfocarse específicamente en la optimización de hiperparámetros.

Variables	Descripción
Prospect ID	Identificador único para cada cliente
Lead Number	Número asignado a cada lead adquirido
Lead Origin	Identificador de la fuente de leads, como API, Landing Page Submission, etc.
Lead Source	Fuente del lead, como Google, búsqueda orgánica, chat de Olark, etc.
Do Not Email	Variable indicadora que muestra si el cliente desea recibir correos electrónicos o no.
Do Not Call	Variable indicadora que muestra si el cliente desea recibir llamadas o no.
Converted	Variable objetivo que indica si el cliente potencial se ha convertido o no.
TotalVisits	Número total de visitas del cliente al sitio web.
Total Time Spent on the Website	Tiempo total pasado por el cliente en el sitio web.
Page Views Per Visit	Número promedio de páginas vistas en cada visita.
Last Activity	Última actividad realizada por el cliente, como Correo electrónico abierto, Conversación de chat de Olark, etc.
Country	País del cliente.
Specialization	Ámbito industrial en el que el cliente trabajaba anteriormente.
How did you hear about X Education	Fuente por la que el cliente conoció X Education.
What is your current occupation	Variable indicadora de si el cliente es estudiante, desempleado o empleado.
What matters most to you in choosing this course	Opción seleccionada por el cliente indicando su motivo principal para realizar el curso.
Search	Variable indicadora de si el cliente vio el anuncio en un motor de búsqueda.
Magazine	Variable indicadora de si el cliente vio el anuncio en una revista.
Newspaper Article	Variable indicadora de si el cliente vio el anuncio en un artículo de periódico.
X Education Forums	Variable indicadora que indica si el cliente vio el anuncio en foros de X Education.
Newspaper	Variable indicadora de si el cliente vio el anuncio en un periódico.
Digital Advertisement	Variable indicadora que indica si el cliente vio el anuncio en un anuncio digital.
Through Recommendations	Variable indicadora de si el cliente llegó a través de recomendaciones.
Receive More Updates About Our Courses	Variable indicadora que muestra si el cliente eligió recibir actualizaciones sobre los cursos o no.
Tags	Etiquetas asignadas a los clientes que indican el estado actual del cliente potencial.
Lead Quality	Variable indicadora de la calidad del lead basada en los datos y la intuición del empleado asignado al lead.
Update me on Supply Chain Content	Variable indicadora que muestra si el cliente desea o no actualizaciones sobre el Contenido de la Cadena de Suministro.
Get updates on DM Content	Variable indicadora que muestra si el cliente desea o no actualizaciones sobre el Contenido DM.
Lead Profile	Nivel del lead asignado a cada cliente en función de su perfil.
City	Ciudad del cliente.
Asymmetrique Activity Index	Índice de Asimetría del Cliente basado en su actividad.
Asymmetrique Profile Index	Índice de Asimetría del Cliente basado en su perfil.
Asymmetrique Activity Score	Puntuación asimétrica del cliente en función de su actividad.
Asymmetrique Profile Score	Puntuación de asimétrica del cliente basado en su perfil.
I agree to pay the amount through cheque	Variable indicadora que muestra si el cliente aceptó o no pagar mediante cheque.
A free copy of Mastering The Interview	Variable indicadora que muestra si el cliente desea o no una copia gratuita ''Mastering the Interview'' o no.
Last Notable Activity	Última actividad destacada realizada por el alumno.

Table 1. Variable y descripción del dataset Lead X education.

Dataset preparation

Antes del entrenamiento de los modelos de ML, se realizó la preparación de los datos. Primero, se filtraron las columnas que no brinden información relevante o que puedan perjudicar el entrenamiento del modelo. Segundo, se remplazaron los valores nulos de las variables categóricas utilizando la moda y la media para las variables numéricas. Tercero, se eliminaron los datos anómalos o ruidosos. Y finalmente, se codificaron las variables categóricas y se guardaron para su uso posterior.

Bayesian Optimization for each algorithm

Para optimizar los hiperparámetros de los cinco algoritmos de aprendizaje automático, estos son, logistic regression, decision tree, random forest, support vector machine y extreme gradient boosting se propone la aplicación de la optimización bayesiana. El objetivo es encontrar la configuración de hiperparámetros que mejoren el rendimiento de los modelos y determinar cual algoritmo optimizado del grupo predice mejor según las métricas de rendimiento. El método de aplicación de esta técnica de optimización comienza con la preparación de los datos y el posterior uso de estos en el entrenamiento y evaluación de los modelos con los hiperparámetros por defecto para determinar las métricas base que se buscará optimizar. Se utilizará el mismo entorno para la ejecución de la optimización bayesiana de cada algoritmo. Este entorno cuenta con los recursos memoria y procesamiento necesarios para realizar la tarea. La computadora utilizada para obtener los resultados cuenta con las siguientes características: procesador Intel Core i7-9750H 2.60Ghz, memoria RAM de 16GB, tarjeta gráfica Nvidia GeForce GTX 1650, almacenamiento SSD de 256GB. Para realizar la optimización primero se definirá un espacio de hiperparámetros y una función objetivo para evaluar calidad del modelo con cierta configuración de hiperparámetros en la predicción de clientes potenciales. La función objetivo se plantea como el promedio del área bajo la curva característica operativa del receptor (ROC) del modelo, calculada mediante validación cruzada de plegado estratificado de tres grupos. Después, se utiliza la optimización bayesiana para mejorar la función objetivo en el espacio de hiperparámetros definido. Se emplea TPE como función sustituta para modelar la función objetivo.

La Tabla 2 muestra los espacios de hiperparámetros en los cuales el algoritmo de optimización bayesiana buscó la configuración óptima para cada uno de los algoritmos. Los mejores hiperparámetros encontrados por la optimización bayesiana serán configurados en los modelos base que serán nuevamente entrenados y evaluados para generar las nuevas métricas de rendimiento. Se utiliza una validación cruzada con plegado estratificado de 3 grupos para evaluar tanto los modelos base como los modelos optimizados, y así comparar sus métricas. El método propuesto está plasmado en la Figura 1.

Modelo	Hiperparámetro	Espacio
RF	criterion	gini or entropy
RF	max_depth	10:200
RF	max_features	sqrt or log2
RF	min_samples_leaf	1:10
RF	n_estimators	200:1000
LR	C	0.05:3
LR	fit_intercept	True or False
LR	max_iter	5:1000
LR	solver	newton-cg, lbfgs, liblinear, sag, saga
LR	tol	1e-05:0.0001
LR	warm_start	True or False
DT	max_depth	10:150
DT	max_leaf_nodes	10:100
DT	min_samples_leaf	1:10
SVM	C	0.01:1000
SVM	gamma	0.0001:0.1
XGB	colsample_bytree	0.1:1
XGB	gamma	0.1:1.3
XGB	learning_rate	0.01:0.5
XGB	max_depth	1:20
XGB	min_child_weight	1:10
XGB	n_estimators	100:1000
XGB	subsample	0.5:1

Table 2. Espacios de hiperparámetros de los modelos de ML utilizados para la Optimización Bayesiana.

Performance Metrics

La evaluación del rendimiento de los modelos de aprendizaje automático en una tarea de clasificación binaria, utilizando un conjunto de datos con un ligero desequilibrio de clases, implicará la aplicación de las siguientes métricas de evaluación.

La métrica accuracy cuantifica el porcentaje general de predicciones correctas y se calcula sumando el número de verdaderos positivos y verdaderos negativos, y dividiéndolo por la suma de verdaderos positivos, verdaderos negativos, falsos negativos y falsos positivos, y puede expresarse como:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Precision cuantifica la proporción de verdaderos positivos en los resultados predichos y se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos positivos, y puede expresarse como:

$$Precision = \frac{TP}{TP + FP}$$

Recall evalúa la proporción de verdaderos positivos correctamente identificados entre todos los verdaderos positivos reales y se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos, y puede expresarse como:

$$Recall = \frac{TP}{TP + FN}$$

F1 score es una métrica compuesta que combina el accuracy y el recall. Se calcula multiplicando dos por el producto de la precisión y la exhaustividad, dividido por la suma de la precisión y la exhaustividad, y puede expresarse como:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

El AUC (Área bajo la Curva ROC) cuantifica el rendimiento general de las predicciones de un modelo, independientemente del umbral de clasificación. Se calcula mediante el cálculo del área bajo la curva ROC. La curva ROC se genera trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) en varios umbrales de clasificación, con

$$TPR = \frac{TP}{TP + FN}$$

y

$$FPR = \frac{TN}{TN + FP}$$

Resultados

En el experimento realizado en este estudio, se utilizó el algoritmo de optimización bayesiana para encontrar la mejor configuración de hiperparámetros de los algoritmos de machine learning RF, LR, DT, SVM y XGBoost. El algoritmo se configuró para realizar 100 evaluaciones, en las cuales se empleó una función de adquisición para identificar la configuración de hiperparámetros que maximizara la función objetivo. La función objetivo se evaluó calculando el promedio de tres métricas de AUC obtenidas a partir de la validación cruzada estratificada de tres partes de cada modelo de aprendizaje automático. La validación cruzada estratificada se realizó utilizando todo el conjunto de datos preparado. El algoritmo de optimización bayesiana se aplicó a cada modelo con una configuración de 100 evaluaciones de la función objetivo. Los hiperparámetros que se optimizaron para cada modelo, junto con el espacio de búsqueda, el valor predeterminado y el mejor valor encontrado para cada hiperparámetro, se describen en la Tabla 3

Modelo	Hiperparámetro	Espacio	Por defecto	Mejor
RF	criterion	gini or entropy	gini	entropy
RF	max_depth	10:200	None	25
RF	max_features	sqrt or log2	sqrt	sqrt
RF	min_samples_leaf	1:10	1	2
RF	n_estimators	200:1000	100	334
LR	C	0.05:3	1.0	0.8741513220219679
LR	fit_intercept	True or False	True	False
LR	max_iter	5:1000	100	903
LR	solver	newton-cg, lbfgs, liblinear, sag, saga	lbfgs	newton-cg
LR	tol	1e-05:0.0001	0.0001	4.8035890572906094e-05
LR	warm_start	True or False	False	True
DT	max_depth	10:150	None	146
DT	max_leaf_nodes	10:100	None	75
DT	min_samples_leaf	1:10	1.0	6
SVM	C	0.01:1000	1.0	338.70008889142264
SVM	gamma	0.0001:0.1	scale	0.00012172594346575304
XGB	colsample_bytree	0.1:1	None	0.7213798117814745
XGB	gamma	0.1:1.3	None	0.5362724143598143
XGB	learning_rate	0.01:0.5	None	0.010781504166149898
XGB	max_depth	1:20	None	6
XGB	min_child_weight	1:10	None	1
XGB	n_estimators	100:1000	100.0	458
XGB	subsample	0.5:1	None	0.7157246177634055

Table 3. Hiperparámetros optimizados de cada modelo usando optimización Bayesiana.

Los algoritmos analizados en el estudio se entrenaron utilizando los mejores valores de los hiperparámetros encontrados mediante la optimización. El rendimiento de cada algoritmo se evaluó utilizando las métricas de precisión, exhaustividad, puntuación F1 y AUC, y se realizó una comparación entre el rendimiento de los algoritmos utilizando sus hiperparámetros base y los hiperparámetros optimizados.

Los resultados mostraron que todos los algoritmos optimizados lograron valores de AUC más altos en comparación con sus versiones base. Entre los algoritmos, Extreme Gradient Boosting tuvo el mejor rendimiento tanto en las métricas base como en las optimizadas, seguido de Random Forest. Si bien el AUC optimizado de Random Forest fue mayor que su resultado base, los demás resultados de métricas base fueron ligeramente más altos que los resultados optimizados. Por otro lado, el Support-Vector Machine con hiperparámetros predeterminados mostró el peor rendimiento. Aunque la versión optimizada mejoró en comparación con sus resultados base, aún quedó rezagada detrás de los otros algoritmos. Los resultados de los

algoritmos base se presentan en la Tabla 4, mientras que los resultados de los algoritmos optimizados se pueden encontrar en la Tabla 5.

Métrica	RF	LR	DT	SVM	XGB
Accuracy	0.918472	0.876310	0.891684	0.727929	0.922199
Precision	0.902523	0.856121	0.856466	0.684385	0.901947
Recall	0.878286	0.809526	0.855061	0.513469	0.890367
F1	0.890087	0.831036	0.855429	0.586709	0.895917
AUC	0.965523	0.938211	0.888183	0.765065	0.972137

Table 4. Métricas de rendimiento de los modelos base.

Métrica	RF	LR	DT	SVM	XGB
Accuracy	0.915793	0.912648	0.896343	0.857675	0.925809
Precision	0.900908	0.896056	0.851956	0.817395	0.905864
Recall	0.872712	0.868997	0.883866	0.801168	0.895945
F1	0.886289	0.882115	0.866138	0.808875	0.900793
AUC	0.969192	0.965441	0.953219	0.931047	0.975943

Table 5. Métricas de rendimiento de los modelos optimizados.

La Figura 2 ilustra la comparación de los porcentajes de precisión entre los modelos optimizados. El algoritmo XGB obtuvo la precisión más alta, con un 92.58%, mientras que el algoritmo SVM tuvo el resultado más bajo del grupo, con un 85.77%. Además, la Figura 3 presenta los resultados de precisión de los modelos optimizados, donde XGB destaca con la precisión más alta, con un 90.59%. En la Figura 4, se puede observar que, entre los algoritmos optimizados, XGB tiene la mejor exhaustividad, con un valor de 89.59%. De manera similar, la Figura 5 muestra las puntuaciones F1, donde el algoritmo optimizado XGB obtuvo la puntuación más alta, con un 90.08% en comparación con los otros algoritmos. Por último, la Figura 6 presenta la comparación de los resultados de la métrica AUC. Aunque todos los algoritmos tuvieron un buen rendimiento en términos de AUC, XGB logró el resultado más alto, con un 97.59%.

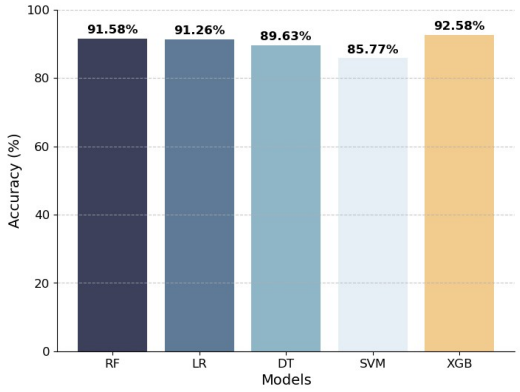

Figure 2. Accuracy de los modelos optimizados.

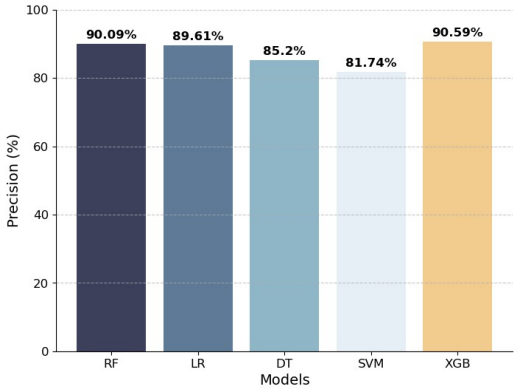

Figure 3. Precision de los modelos optimizados.

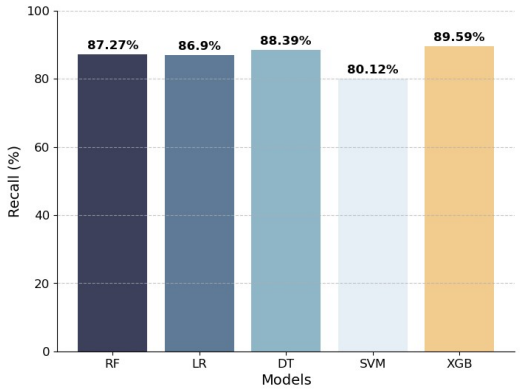

Figure 4. Recall de los modelos optimizados.

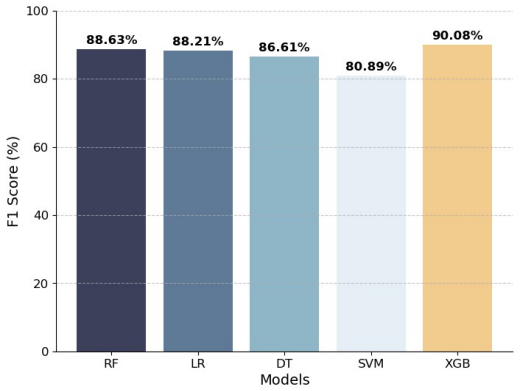

Figure 5. F1 Score de los modelos optimizados.

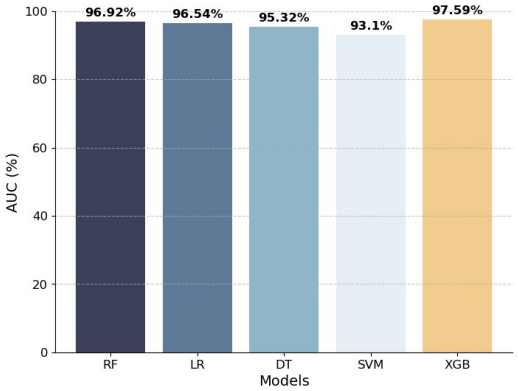

Figure 6. AUC de los modelos optimizados.

La duración de 100 evaluaciones del algoritmo de optimización bayesiana fue medida para cada algoritmo de aprendizaje automático. Entre los algoritmos, SVM tuvo el tiempo de ejecución más largo, con 1717 segundos, mientras que DT tuvo el tiempo de ejecución más corto, con 18 segundos. La Figura 7 muestra una comparación de la duración de la optimización bayesiana de hiperparámetros para cada uno de los algoritmos utilizados.

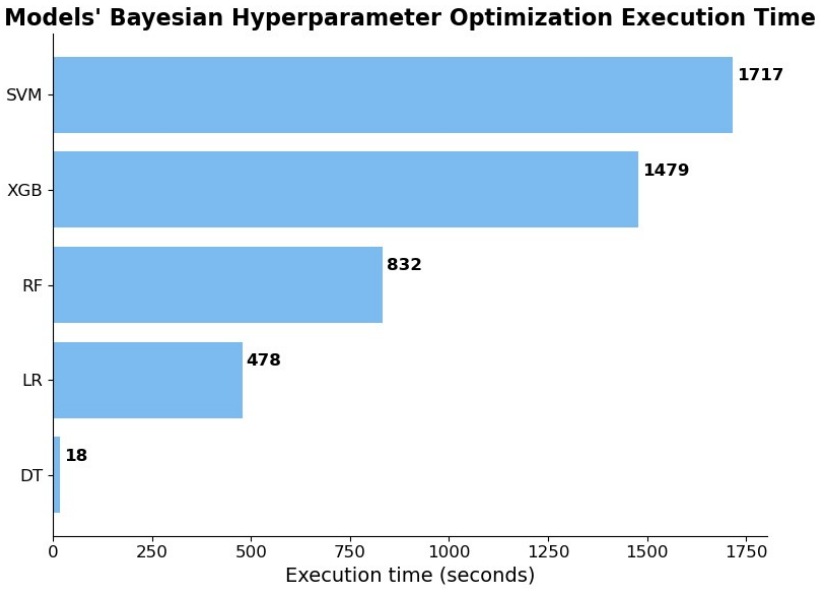

Figure 7. Comparación del tiempo de ejecución de la optimización bayesiana de hiperparámetros para todos los modelos.

Observaciones finales

En conclusión, nuestro estudio se centró en la aplicación de algoritmos de aprendizaje automático para la puntuación de clientes potenciales en marketing, con el objetivo de optimizar la conversión de clientes en el ámbito del e-learning. Abordamos el desafío de la optimización de parámetros, que tiene un impacto significativo en el rendimiento de estos modelos.

Aprovechando el poder de la optimización bayesiana, introdujimos un enfoque novedoso para el ajuste de hiperparámetros, que explora de manera eficiente el espacio de hiperparámetros de alta dimensionalidad y encuentra configuraciones óptimas para los algoritmos de aprendizaje automático. Nuestros resultados demostraron la efectividad de la optimización bayesiana de hiperparámetros en la mejora del rendimiento de los modelos de predicción de conversión de clientes en e-learning, superando el uso de valores de hiperparámetros predeterminados.

Entre los algoritmos analizados, el algoritmo Extreme Gradient Boosting (XGB) superó consistentemente a los demás, mostrando los valores más altos en precisión, exhaustividad, puntuación F1 y AUC. Exhibió un rendimiento excepcional tanto en las métricas base como en las optimizadas, destacando su efectividad en la puntuación de clientes potenciales y la predicción de conversión. Por otro lado, el algoritmo Support Vector Machine (SVM) mostró el peor rendimiento, incluso después de la optimización de hiperparámetros. Si bien la versión optimizada demostró algunas mejoras en comparación con los resultados base, aún quedó rezagada detrás de los demás algoritmos en términos de rendimiento general. En cuanto al

tiempo de ejecución, SVM tuvo la duración más larga para la optimización bayesiana de hiperparámetros, mientras que Decision Tree (DT) tuvo la más corta.

Los hallazgos de nuestro estudio resaltan la importancia de la optimización de hiperparámetros para lograr un rendimiento óptimo de los modelos de aprendizaje automático. Al aprovechar la optimización bayesiana, exploramos de manera efectiva el espacio de parámetros e identificamos configuraciones superiores que condujeron a una mejor puntuación de clientes potenciales y predicción de conversión.

Nuestra investigación contribuye al creciente cuerpo de conocimiento en el campo de la toma de decisiones basada en datos y destaca la importancia de incorporar técnicas avanzadas, como la optimización bayesiana, en las estrategias de marketing. Los algoritmos optimizados ofrecen a los especialistas en marketing ideas y herramientas valiosas para priorizar esfuerzos, asignar recursos de manera eficiente y llevar a cabo campañas de marketing exitosas.

En futuras investigaciones, sería valioso explorar la aplicación de la optimización bayesiana de hiperparámetros en otros ámbitos y ampliar el análisis para incluir algoritmos adicionales de aprendizaje automático. Además, investigar el impacto de los algoritmos optimizados en las tasas de conversión y las métricas de participación de los clientes proporcionaría más evidencia de su eficacia práctica.

En resumen, nuestro estudio demuestra que la optimización bayesiana de hiperparámetros mejora el rendimiento de los algoritmos de aprendizaje automático en la puntuación de clientes potenciales y la predicción de conversión de clientes. Al aprovechar los algoritmos optimizados, las organizaciones pueden tomar decisiones basadas en datos, maximizar las tasas de conversión y lograr mejores resultados comerciales en la industria del e-learning.

Disponibilidad de los datos

El dataset “Lead Scoring X Online Education,” está disponible de forma gratuita en el repositorio Kaggle:
https://www.kaggle.com/datasets/lakshmikalyan/lead-scoring-x-online-education

Referencias

1. Duncan, B. A. & Elkan, C. P. Probabilistic modeling of a sales funnel to prioritize leads. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1751–1758 (2015).
2. Terho, H., Mero, J., Siutla, L. & Jaakkola, E. Digital content marketing in business markets: Activities, consequences, and contingencies along the customer journey. *Ind. Mark. Manag.* 105, 294–310 (2022).
3. D’Haen, J. & Van den Poel, D. Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. *Ind. Mark. Manag.* 42, 544–551 (2013).
4. Lindahl, E. A qualitative examination of lead scoring in b2b marketing automation, with a recommendation for its practice. (2017).
5. Wamba-Taguimdje, S.-L., Fosso Wamba, S., Kala Kamdjoug, J. R. & Tchatchouang Wanko, C. E. Influence of artificial intelligence (ai) on firm performance: the business value of ai-based transformation projects. *Bus. Process. Manag. J.* 26, 1893–1924 (2020).
6. Davenport, T. H. From analytics to artificial intelligence. *J. Bus. Anal.* 1, 73–80 (2018).
7. Miklosik, A., Kuchta, M., Evans, N. & Zak, S. Towards the adoption of machine learning-based analytical tools in digital marketing. *Ieee Access* 7, 85705–85718 (2019).

8. Overgoor, G., Chica, M., Rand, W. & Weishampel, A. Letting the computers take over: Using ai to solve marketing problems. *California Manag. Rev.* 61, 156–185 (2019).

9. Yuan, K.-C. *et al.* Using transfer learning method to develop an artificial intelligence assisted triaging for endotracheal tube position on chest x-ray. *Diagnostics* 11, 1844 (2021).

10. Wang, A., Xu, J., Tu, R., Saleh, M. & Hatzopoulou, M. Potential of machine learning for prediction of traffic related air pollution. *Transp. Res. Part D: Transp. Environ.* 88, 102599 (2020).

11. Hutter, F., Kotthoff, L. & Vanschoren, J. *Automated machine learning: methods, systems, challenges* (Springer Nature, 2019).

12. Zöller, M.-A. & Huber, M. F. Benchmark and survey of automated machine learning frameworks. *J. artificial intelligence research* 70, 409–472 (2021).

13. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Adv. neural information processing systems* 24 (2011).

14. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. machine learning research* 13 (2012).

15. Boelrijk, J., Pirok, B., Ensing, B. & Forré, P. Bayesian optimization of comprehensive two-dimensional liquid chromatography separations. *J. Chromatogr. A* 1659, 462628 (2021).

16. Calandra, R., Seyfarth, A., Peters, J. & Deisenroth, M. P. An experimental comparison of bayesian optimization for bipedal locomotion. In *2014 IEEE international conference on robotics and automation (ICRA)*, 1951–1958 (IEEE, 2014).

17. Victoria, A. H. & Maragatham, G. Automatic tuning of hyperparameters using bayesian optimization. *Evol. Syst.* 12, 217–223 (2021).

18. Wu, J. *et al.* Hyperparameter optimization for machine learning models based on bayesian optimization. *J. Electron. Sci. Technol.* 17, 26–40 (2019).

19. Saufi, S. R., Ahmad, Z. A. B., Leong, M. S. & Lim, M. H. Gearbox fault diagnosis using a deep learning model with limited data sample. *IEEE Transactions on Ind. Informatics* 16, 6263–6271 (2020).

20. Attia, P., Deetjen, M. & Witmer, J. Accelerating battery development via early prediction of cell lifetime. *Elastic* 2, 2 (2018).

21. Delen, D. & Ram, S. Research challenges and opportunities in business analytics. *J. Bus. Anal.* 1, 2–12 (2018).

22. Robert, N. & József, M. Automating lead scoring with machine learning: An experimental study. DOI: 10.24251/HICSS. 2020.177 (2020).

23. Ma, J., Zhang, J., Li, R., Zheng, H. & Li, W. Using bayesian optimization to automate the calibration of complex hydrological models: Framework and application. *Environ. Model. & Softw.* 147, 105235 (2022).

24. Niedermayer, D. An introduction to bayesian networks and their contemporary applications. *Innov. Bayesian networks: Theory applications* 117–130 (2008).

25. Pourmohamad, T. & Lee, H. K. Bayesian optimization via barrier functions. *J. Comput. Graph. Stat.* 31, 74–83 (2022).

26. Feurer, M. & Hutter, F. Hyperparameter optimization. *Autom. machine learning: Methods, systems, challenges* 3–33 (2019).

27. Jo, Y., Min, K., Jung, D., Sunwoo, M. & Han, M. Comparative study of the artificial neural network with three hyperparameter optimization methods for the precise lp-egr estimation using in-cylinder pressure in a turbocharged gdi engine. *Appl. Therm. Eng.* 149, 1324–1334 (2019).

28. Charbuty, B. & Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* 2, 20–28 (2021).

29. Pandimurugan, V., Usha, D., Guptha, M. N., Hema, M. *et al.* Random forest tree classification algorithm for predicating loan. *Mater. Today: Proc.* 57, 2216–2222 (2022).

30. Shipe, M. E., Deppen, S. A., Farjah, F. & Grogan, E. L. Developing prediction models for clinical use using logistic regression: an overview. *J. thoracic disease* 11, S574 (2019).

31. Hafdaoui, H., Chahtou, A., Bouchakour, S., Belhaouas, N. *et al.* Analyzing the performance of photovoltaic systems using support vector machine classifier. *Sustain. Energy, Grids Networks* 29, 100592 (2022).

32. Sibindi, R., Mwangi, R. W. & Waititu, A. G. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. *Eng. Reports* e12599 (2022).

33. Kalyan, L. Lead scoring x online education. *kaggle* https://www.kaggle.com/datasets/lakshmikalyan/ lead-scoring-x-online-education (2019).

34. Jadli, A., Hamim, M., Hain, M. & Hasbaoui, A. Toward a smart lead scoring system using machine learning. *Indian J. Comput. Sci. Eng.* 13, 433 – 443 (2022).

ANEXOS

Scientific Reports - Receipt of Manuscript 'Bayesian optimization of...'

Scientific Reports <srep@nature.com>

Mié 05/07/2023 21:35

Para:Jorge Maquera Canales <jorgemaquerac@upeu.edu.pe>

Ref: Submission ID 47e926ff-a1da-4afc-bc4f-e7a8870b2541

Dear Dr Maquera,

Please note that you are listed as a co-author on the manuscript "Bayesian optimization of machine learning models to improve e-learning customer prediction", which was submitted to Scientific Reports on 06 July 2023 UTC.

If you have any queries related to this manuscript please contact the corresponding author, who is solely responsible for communicating with the journal.

Kind regards,

Peer Review Advisors
Scientific Reports

“AÑO DE LA UNIDAD, LA PAZ Y EL DESARROLLO”

RESOLUCIÓN N° 0088-2023/UPeU-FIA-CF-T

Lima, Ñaña 14 de marzo de 2023

VISTO:

El expediente de **Maquera Canales Jorge Daniel**, identificado(a) con Código Universitario Nº 201810763, de la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión;

CONSIDERANDO:

Que la Universidad Peruana Unión tiene autonomía académica, administrativa y normativa, dentro del ámbito establecido por la Ley Universitaria Nº 30220 y el Estatuto de la Universidad;

Que la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, mediante sus reglamentos académicos y administrativos, ha establecido las formas y procedimientos para la aprobación e inscripción del perfil de proyecto de tesis en formato artículo y la designación o nombramiento del asesor para la obtención del título profesional;

Que **Maquera Canales Jorge Daniel**, ha solicitado: la inscripción del perfil de proyecto de tesis titulado "Optimización bayesiana de modelos de aprendizaje automático para mejorar la predicción de clientes e-learning" y la designación del Asesor, encargado de orientar y asesorar la ejecución del perfil de proyecto de tesis en formato artículo;

Estando a lo acordado en la sesión del Consejo de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, celebrada el 14 de marzo de 2023, y en aplicación del Estatuto y el Reglamento General de Investigación de la Universidad;

SE RESUELVE:

Aprobar el perfil de proyecto de tesis en formato artículo titulado **"Optimización bayesiana de modelos de aprendizaje automático para mejorar la predicción de clientes e-learning"** y disponer su inscripción en el registro correspondiente, designar como asesor a **Mg. Saboya Rios Nemias para** que oriente y asesore la ejecución del perfil de proyecto de tesis en formato artículo el cual fue dictaminado por: **Ph.D. Javier Linkolk López Gonzales y MSc. Fredy Abel Huanca Torres**, otorgándoles un plazo máximo de doce (12) meses para la ejecución.

Regístrese, comuníquese y archívese.

Dra. Erika Inés Acuña Salinas
DECANA

Dr. Santiago Ramírez López
SECRETARIO ACADÉMICO

cc:
-Interesado
-Asesor
-Dirección General de Investigación
-Archivo