

UNIVERSIDAD PERUANA UNIÓN
FACULTAD DE INGENIERIA Y ARQUITECTURA
Escuela Profesional de Ingeniería de Sistemas



**Machine learning con boosting sobre la percepción del servicio
de Internet en Perú**

Tesis para obtener el Título Profesional de Ingeniero de Sistemas

Autor:

Bach. Alejandro Luis Javier Gutierrez Beltran
Bach. Arnold Fernandez Blanco

Asesor:

Dr. Juan Jesús Soria Quijaite

Lima, 14 de Julio de 2025

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Dr. Soria Quijaite Juan Jesus, docente de la Facultad de Ingeniería y Arquitectura, de la Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **“Machine learning con boosting sobre la percepción del servicio de Internet en Perú”** de los autores Gutierrez Beltran, Alejandro Luis Javier y Fernandez Blanco, Arnold tiene un índice de similitud de 5 % 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 14 días del mes de julio del año 2025.

A handwritten signature in black ink, appearing to read 'Juan Soria', with a long horizontal line extending to the right.

Dr. Soria Quijaite Juan Jesus

000

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Naña, Villa Unión, a los **08** día(s) del mes de **julio** del año 2025 siendo **las 09:00 horas**, se reunieron en modalidad virtual u online sincrónica, bajo la dirección del Señor Presidente del jurado: **Mg. Immer Elias Cuellar Rodriguez**, el secretario: **Mg. Lennin Henry Centurion Julca**, y los demás miembros: **Mg. Nemias Saboya Rios y el Mg. Ferdinan Edgardo Pineda Anco**, y el asesor, **Dr. Juan Jesús Soria Quijaite**, con el propósito de administrar el acto académico de sustentación de la tesis titulada: "Machine learning con boosting sobre la percepción del servicio de Internet en Perú"

de el(los)/la(las) bachiller/es: a) **Arnold Fernández Blanco**

..... b) **Alejandro Luis Javier Gutiérrez Beltrán**

.....

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): **ARNOLD FERNÁNDEZ BLANCO**

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
APROBADO	18	A-	MUY BUENO	SOBRESALIENTE

Candidato (b): **ALEJANDRO LUIS JAVIER GUTIÉRREZ BELTRÁN**

CALIFICACIÓN	ESCALAS			
	Vigesimal	Literal	Cualitativa	
APROBADO	18	A-	MUY BUENO	SOBRESALIENTE

(*) 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
Mg. Immer Elias
Cuellar Rodriguez



Secretario
Mg. Lennin Henry
Centurion Julca

Asesor
Dr. Juan Jesús Soria
Quijaite

Miembro
Mg. Nemias Saboya
Rios

Miembro
Mg. Ferdinan Edgardo
Pineda Anco

Candidato/a (a)
Arnold Fernández
Blanco

Candidato/a (b)
Alejandro Luis Javier
Gutiérrez Beltrán

DEDICATORIA

Agradezco profundamente a Dios, por ser
mi guía en este camino.

A mis padres, Epifanio y Edi, por creer
siempre en mí y darme esta oportunidad.
A mi hermana, por su apoyo
incondicional y a Mini, por su compañía
silenciosa que me brindó en los días
difíciles de estudio.

Con cariño Arnold F.B

Agradezco a Dios y a mis padres Luis
Gutierrez y Mery Beltran por haberme
brindado esta oportunidad tan maravillosa
de ser profesional.

De Alejandro G.B

ÍNDICE

RESUMEN.....	6
ABSTRACT	6
1. INTRODUCCION	7
2. METODOLOGIA.....	9
2.1. Dataset	10
2.2. Limpieza de datos.....	10
2.3. Selección de parámetros relevantes	10
2.4. Fase de Modelado	10
2.4.1. Árbol de Decisión	11
2.4.2. Extreme Gradient Boosting (XGBoost)	12
2.4.3. Random Forest	13
2.4.4. Regresión Logística	14
2.4.5. K-Nearest Neighbors	14
2.5. Construcción del Metamodelo	15
2.5.1. Matriz de confusión:	16
3. RESULTADOS Y DISCUSIÓN	17
3.1. Análisis descriptivo de la percepción del usuario	17
3.2. Resultado de la clasificación de patrones	18
3.3. Resultados de la clasificación de la percepción del servicio de Internet (comparación de métricas).....	20
3.4. Comparación del modelo Ensamble Boosting y Modelos Individuales.....	21
4. CONCLUSIÓN.....	22
5. REFERENCIAS	23

Machine learning con boosting sobre la percepción del servicio de Internet en Perú

RESUMEN

El acceso a Internet de calidad es esencial para el desarrollo social y económico, pero el Perú aún tiene grandes brechas en el conocimiento de sus servicios. Este estudio propone un metamodelo de aprendizaje automático basado en el aumento de la integración para clasificar los patrones de calidad percibida de servicio de Internet residencial en Perú. Se preprocesaron 129,352 registros utilizando datos del INEI utilizando SMOTE para equilibrar categorías y seleccionar variables claves relacionadas con satisfacción, calidad, demora y servicio al cliente. Los modelos utilizados incluyen XGBoost, LightGBM, Random Forest y Logistic Regresion. Los resultados muestran que el metamodelo propuesto mejora significativamente la precisión en comparación con un modelo único con un AUC promedio al 80%. Este enfoque no sólo identifica patrones de percepción de los usuarios recomendaciones prácticas para los operadores de telecomunicaciones, enfatizando la necesidad de optimizar la infraestructura y reducir la brecha digital del país.

Palabras clave: Calidad del servicio de Internet, Percepción del usuario, Aprendizaje automático, Metamodelo de clasificación.

ABSTRACT

Quality Internet access is essential for social and economic development, but Peru still has significant gaps in knowledge of its services. This study proposes a machine learning metamodel based on increased integration to classify the patterns of perceived quality of residential Internet service in Peru. One hundred twenty-nine thousand three hundred fifty-two records were pre-processed using INEI data and SMOTE to balance categories and select key variables related to satisfaction, quality, delay, and customer service. Models used include XGBoost, LightGBM, Random Forest, and Logistic Regression. The results show that the proposed metamodel significantly improves accuracy compared to a single model with an average AUC of 80%. This approach not only identifies patterns of user perception and practical recommendations for telecommunications operators but emphasizes the need to optimize infrastructure and reduce the country's digital divide.

Keywords: Internet service quality, User perception, Machine learning, Classification metamodel.

1. INTRODUCCION

El internet se ha convertido en una necesidad fundamental en los últimos años, permitiendo la conexión entre diversos dispositivos y sistemas, desempeñando un papel crucial en el ámbito educativo [1]. No solo facilita el acceso a información y recursos globales, sino que también mejora la calidad de la enseñanza educativa, superando barreras geográficas al ofrecer acceso a zonas remotas, posibilitando la enseñanza a distancia y el monitoreo del desempeño estudiantil [2]. Además, fomenta entornos de aprendizaje inclusivos y adaptativos, promoviendo la colaboración global, la participación en experiencias educativas interactivas, el desarrollo personal y profesional de los estudiantes [3], [4]. Su capacidad para integrar tecnología en la educación impulsa la innovación y prepara al entorno estudiantil para enfrentar los desafíos de un mundo digitalizado [5]. Sin embargo, en América Latina, el acceso a internet ha revelado disparidades económicas y sociales, evidenciando una brecha considerable entre las zonas rurales y urbanas [6], [7]. En el estudio [8] se investiga la relación entre el acceso a Internet y la innovación en 29 países, encontrando que el acceso a la red tiene un efecto positivo en la innovación. Sin embargo, el estudio resalta que no solo el acceso es determinante, sino también la calidad del servicio, incluyendo velocidad, estabilidad y disponibilidad, factores clave para potenciar la innovación. Por otro lado, en los Estados Unidos se realizó una encuesta a 5,200 personas sobre la calidad y velocidad del servicio de Internet, los resultados revelaron que los usuarios valoran más la estabilidad y confiabilidad de la conexión que la velocidad, considerándolas esenciales para una experiencia óptima [9]. En el caso de Perú, con la llegada de la pandemia de COVID-19, las universidades cerraron las aulas presenciales y se adoptó la educación en línea. Al inicio del semestre, el 51.1% de los estudiantes tenía una percepción negativa sobre este modelo educativo, aunque al final del ciclo la actitud mejoró ligeramente al 48.1%. Entre los factores que influyeron en esta percepción destacan la conectividad y la calidad del servicio de Internet [10]. El acceso a internet en las instituciones educativas según la UGEL (Unidades de Gestión Educativa Local), es crucial para mejorar la eficiencia del sistema y reducir las brechas de aprendizajes en los estudiantes peruanos [11]. Sin embargo, un estudio realizado en niños de escuelas públicas en Lima encontró que, si bien los estudiantes demostraron un mayor dominio en el uso de computadoras, no se evidenciaron mejoras significativas en su rendimiento académico. Este estudio señala que el acceso a Internet, por sí solo, no es suficiente para mejorar el aprendizaje si la calidad del servicio no está alineada con un uso educativo adecuado [12]. A nivel nacional, más del 60% de los peruanos no cuentan con acceso a Internet, la calidad del servicio de internet enfrenta serios desafíos, incluyendo baja velocidad, inestabilidad en la conexión y cobertura insuficiente, lo que genera una percepción generalizada de insatisfacción entre los usuarios. En las zonas rurales, solo el 10% de los hogares tiene acceso a Internet, frente a un promedio nacional del 67%, según cifras del Ospitel, esta brecha no solo limita el acceso a la información, sino que también impacta negativamente en el desarrollo económico [13]. En el departamento de Puno, se ha observado que contar con un servicio de calidad puede resultar en un aumento de ingresos de hasta un 45% [14].

Para mejorar la percepción de la calidad del internet, los usuarios valoran especialmente la velocidad adecuada y la consistencia del servicio, por otro lado, la insatisfacción en áreas con acceso limitado o deficiente es un problema crítico que requiere atención [15].

Las consecuencias de un mal servicio de internet trascienden la insatisfacción individual, afectando la productividad y el desarrollo económico en general, según estudios internacionales han demostrado que los usuarios están dispuestos a pagar más por mejoras en la calidad de su conexión, lo que refleja una alta demanda por un servicio más estable y rápido, especialmente en regiones con infraestructura deficiente [16]. En este contexto, el abandono de clientes se convierte en un desafío para las empresas de telecomunicaciones, que deben no solo atraer nuevos usuarios, sino también fidelizar a los existentes. La implementación de algoritmos de aprendizaje automático, como Random Forest y XGBoost, es fundamental para identificar patrones en los datos históricos de los clientes y determinar los factores que influyen en su decisión de abandonar el servicio [17], [18]. Al combinar las fortalezas de varios algoritmos mediante metodologías de ensamble, se puede mejorar la clasificación de patrones y, en última instancia, optimizar la experiencia del usuario y garantizar un servicio de internet de calidad [19].

En este contexto, el aprendizaje automático ha tomado bastante respaldo por parte de investigadores y de la comunidad científica alrededor del mundo que intentan sacarle provecho a estas diferentes tecnologías para poder tener una fuerte y sólida base de resultados que aporten al desarrollo de la sociedad [20]. El análisis de rotación de clientes en la industria de telecomunicaciones, donde se empleó un dataset de kaggle para clasificar clientes entre los que abandonan y permanecen, destacan la aplicación de XGBoost como método principal, dicho modelo alcanzó un AUC de 0.65% evidenciando un desempeño moderado con oportunidades claras de optimización [21]. También se han hecho estudios sobre la clasificación de mensajes de texto para identificar si es spam y no spam, utilizando algoritmos como Regresión Logística, SVM, Gradient Boosting y Random Forest, obteniendo métricas significativas como precisión, recall, F1-score, con un incremento del 48.79% en Boosting [22]. Por su parte, la técnica de generación de data sintética SMOTE en conjunto con algoritmos Random Forest y Support Vector Machine (SVM) incrementó el AUC promedio de 61.2% a 82.2%, demostrando así la efectividad de estas técnicas para mejorar la representatividad de clases minoritarias y así el problema de datasets desbalanceados [23]. Continuando ahora con el ámbito del mundo de las noticias, el modelo de LightGBM alcanzó un AUC casi perfecto de 0.999, consolidándose, así como una de las mejores opciones, eficiente y mejora frente a otros modelos más complejos [24]. Así finalmente, otra investigación enfocada a la predicción de enfermedades cardíacas, pero ahora empleando la técnica de ensamble basado en boosting para mejorar el rendimiento en problemas de clasificación complejos, alcanzó un AUC superiores al 93% en diferentes datasets, demostrando así la capacidad de los métodos de boosting para optimizar el rendimiento de estos mismos [25]. Dichos resultados mencionados anteriormente destacan cómo la métrica del AUC es esencial no solo como criterio de evaluación, sino también como una herramienta estratégica para guiar la elección y construcción de modelos en estudios de clasificación, como el análisis del servicio de calidad del internet peruano.

La investigación, desarrolla un metamodelo basado en técnicas de ensamble boosting, utilizando algoritmos como XGBoost y LightGBM, esto con el objetivo de clasificar patrones en la percepción de la calidad del servicio de internet en hogares peruanos. Dicho enfoque requiere no solo de mejorar la precisión de las predicciones, sino que también busca optimizar métricas clave como el Área Bajo la Curva (AUC), proporcionando así resultados más confiables y representativos. Para el tema de desbalanceo de clases debido a los parámetros relevantes escogidos para la

investigación, integramos la técnica de Synthetic Minority Oversampling Technique (SMOTE) [26], que garantiza una mayor equidad en la representación de las clases y potencia del rendimiento del modelo. Además de esto, nuestro enfoque añade procesos de selección de características y normalización, asegurando así la robustez del modelo ante datos heterogéneos y variados. El metamodelo tiene como finalidad proporcionar un variado análisis integral y eficiente, adaptables en diferentes contextos relacionados con la calidad del servicio de Internet en el Perú.

El estudio integra técnicas avanzadas de ensamble boosting con algoritmos ya previamente mencionados aplicados a un contexto regional específico como el mercado peruano. Además de esto se aplica la técnica de SMOTE para el balanceo de datos, lo cuál es un problema muy frecuente en datasets de altos volúmenes. La relevancia en esta investigación radica en su capacidad para generar recomendaciones prácticas que orienten a los proveedores de servicios hacia una mejora efectiva de la calidad del servicio de internet.

2. METODOLOGIA

La metodología tiene información sobre los métodos, modelos, materiales y sobre el conjunto de datos que se utilizaron para esta investigación. Las fases más relevantes fueron la limpieza y normalización del dataset, el balanceo de este mismo, el entrenamiento con los modelos y metamodelo y por último la fase de resultados. En la Figura 1 titulada “Boosting Framework for Classification” se presenta el flujo con las 4 fases importantes mencionadas anteriormente.

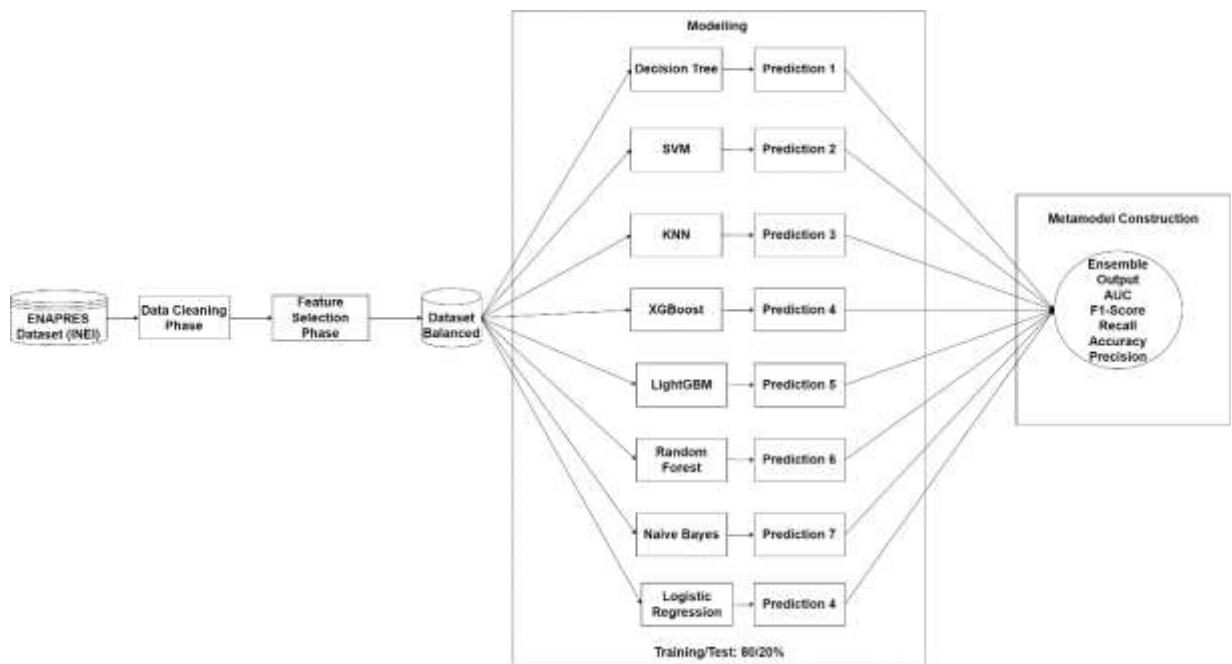


Figura 1. Boosting Framework for Classification

2.1. Dataset

La investigación utilizó datos recolectados de la Encuesta Nacional de Programas Presupuestales 2023 del Instituto Nacional de Estadística e Informática (INEI), que incluye 129,352 registros de hogares peruanos que acceden al servicio de internet. La población abarca tanto regiones urbanas como rurales, permitiendo analizar la diversidad en el acceso y percepción de la calidad del servicio de Internet en todo el país [27].

La muestra seleccionada para el análisis es un subconjunto representativo de la población total, elegido mediante un muestreo aleatorio simple y mediante la fórmula para una población finita [28] que ascienden 70,233 registros de usuarios encuestados. Esto garantiza que todos los hogares tengan igualdad de probabilidades de ser incluidos, asegurando la validez y generalización de los resultados.

2.2. Limpieza de datos

Para la etapa de limpieza de datos se utilizó el método CRISP DM (Cross Industry Standard Process for data Mining) para garantizar la calidad y coherencia de los datos. La metodología de CRISP DM es un marco para la minería de datos que organiza el proceso en etapas claras, comienza desde la comprensión del negocio y de los datos, continúa con el modelado, evaluación y despliegue, y es compatible con diferentes enfoques de calidad y gestión de riesgos [29], [30]. Durante este proceso, los valores faltantes se imputan utilizando la mediana, lo que preserva la integridad de los datos sin introducir sesgos significativos.

2.3. Selección de parámetros relevantes

En esta etapa se tiene el conjunto de datos proporcionado por el Instituto Nacional de Estadística e Informática (INEI), que contiene información relevante sobre el servicio de internet en hogares peruanos. El análisis se enfocó en variables seleccionadas específicamente para abordar los objetivos específicos donde el servicio al cliente relaciona aspectos de la atención brindada a los usuarios en la calidad de soporte del servicio, así mismo la calidad del servicio mide la percepción de los usuarios respecto al desempeño del servicio de Internet. Donde la satisfacción del servicio de internet es un punto importante para la experiencia brindada al cliente, donde el servicio de latencia es primordial en el rendimiento del servicio en los hogares peruanos [31], [32].

El análisis y segmentación de estas variables permiten evaluar múltiples dimensiones de la calidad del servicio de internet en hogares peruanos, estableciendo una relación directa con los objetivos planteados y facilitando la implementación de modelos de clasificación para optimizar los resultados.

2.4. Fase de Modelado

En esta etapa se procedió a la construcción y evaluación de los modelos de clasificación seleccionados para abordar la percepción de la calidad del servicio de internet integrando los modelos de boosting como lo fue XGBoost y

LightGBM, además se utilizaron los modelos como SVM, KNN, NB, Random Forest, Regresión Logística y Árbol de decisión. Con el objetivo de entrenar diferentes algoritmos de machine learning, se utilizó el conjunto de datos balanceado y características previamente seleccionadas. También se utilizó el método de Stacking para poder construir el meta-modelo basado en algoritmos de boosting previamente mencionado y realizar una comparación de los resultados en relación con los otros modelos. Esta etapa es importante, debido a que permite comparar el desempeño de múltiples modelos individuales y sentar las bases para la construcción del metamodelo final. Teniendo en cuenta que en los diferentes modelos se aplicó Grid Search o Random Search según sea el caso para poder afinar los hiperparámetros y mejorar el desempeño.

2.4.1. Árbol de Decisión

El modelo de árboles de decisión está basado en reglas de tipo if-then para predecir una variable, a partir de pruebas booleanas, dicha técnica es sencilla de interpretar, esto debido a su estructura jerárquica que contiene. El algoritmo construye un modelo de entrenamiento aprendiendo reglas a partir de datos previos, para que luego empiece a predecir la clase o valor de nuevas observaciones, donde cada instancia es clasificada desde el nodo raíz hasta alcanzar el nodo hoja que nos muestra la predicción final [33]. Como resultado del modelo en la cuadrícula se utilizó los parámetros: `estimator=dt_model`, `param_distributions=param_grid`, `scoring='f1_weighted'`, `cv=5`, `n_iter=50`, `verbose=1`, `random_state=42`, `n_jobs=-1` para el algoritmo. La fórmula del presente modelo se observa en la ecuación (1):

$$División(S) = \operatorname{argmax}_A \text{ Criterio}(A)$$

Donde:

- *A*: Conjunto de atributos
- *Criterio*: Ganancia de Información, Índice Gini u otra métrica.

Support Vector Machine (SVM)

Es un modelo que utiliza clasificadores que funcionan construyendo hiperplanos en un espacio de características de alta dimensión que intenta maximizar la separación entre diferentes clases. Dicho enfoque hace que las máquinas de vectores de soporte sean efectivas para problemas de clasificación tanto en línea como no lineales, especialmente cuando se utilizan funciones del núcleo para transformar los datos en un espacio separable por clases [34]. Como resultado de la búsqueda de hiperparámetros con `RandomizedSearchCV` se utilizó los parámetros: `estimator=svm_model`, `param_distributions=param_distributions`, `n_iter=10`, `scoring='f1_weighted'`, `cv=3`, `verbose=1`, `n_jobs=-1`, `random_state=42` para el algoritmo SVM. La fórmula del presente modelo se observa en la ecuación (2):

$$\max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i x_j) \quad (2)$$

Donde:

- α_i : Multiplicadores de Lagrange
- y_i : Etiquetas de las clases (-1 o +1)
- $K(x_i x_j)$: Función núcleo que mide la similitud entre las muestras x_i y x_j
- $x_i x_j$: Muestra de datos.
- n : Número total de muestras

2.4.2. Extreme Gradient Boosting (XGBoost)

Es un modelo que se basa en ensambles que usa árboles de decisión como clasificadores para poder optimizar la precisión predictiva mediante técnicas avanzadas como la expresión Taylor para minimizar la función de costo. A comparación de otros métodos de boosting, el presente modelo añade derivadas de segundo orden para ajustar los parámetros del modelo de manera eficiente, lo que permite manejar problemas de sobreajuste y mejorar la generalización [35]. Como resultado de la búsqueda de hiperparámetros se utilizó los parámetros: estimator=xgb_model, param_distributions=param_grid, scoring='f1_weighted', cv=5, n_iter=50, verbose=1, n_jobs=-1 para el algoritmo de XGBoost. La fórmula del presente modelo se muestra en la ecuación (3):

$$L^{(t)} = \sum_{i=1}^n [g_i f_i(x_i) + \frac{1}{2} h_i(x_i)^2] + \Omega(f_t) \quad (3)$$

Donde

- g_i : Primera derivada (gradiente)
- h_i : Segunda derivada (hessiana)
- $f_i(x_i)$: Predicción del modelo en la iteración t

Light gradient boosted machine algorithm

Está basado en boosting de gradiente que usa árboles de decisión como modelos base para tareas de clasificación y regresión. Dicho modelo incorpora técnicas como la de Exclusive Bundling (EFB) y Gradient-based One-Side Sampling (GOSS) para mejorar y optimizar el procesamiento y clasificación de datos, reduciendo así el uso de memoria y acelerando el entrenamiento[36].

Como resultado del modelo de hiperparámetros se obtuvo los parámetros: estimator=lgbm_model, param_distributions=param_distributions, n_iter=

20, scoring='f1_weighted', cv=3, verbose=1, n_jobs=-1, random_state=42 para el algoritmo de LightGBM. La fórmula del presente modelo se muestra en la ecuación (4):

$$L^{(t)} = \sum_{i=1}^n [g_i f_i(x_i) + \frac{1}{2} h_i(x_i)^2] + \Omega(f_t) \quad (4)$$

- Donde
- g_i : Primera derivada (gradiente)
- h_i : Segunda derivada (hessiana)
- $f_i(x_i)$: Predicción del modelo en la iteración t

2.4.3. Random Forest

Es un método de clasificación que está basado en múltiples árboles de decisión a partir de muestras Bootstrap. Cada árbol emite un voto y la clasificación final se determina por mayoría, lo que permite manejar grandes conjuntos de datos desequilibrados. Añadido a esto, se utiliza para la selección de característica mediante el método de Borut para estimar la importancia de las variables en el modelo. Su principal ventaja es su capacidad para reducir el sobreajuste, identificar variables clave y mantener una alta precisión de clasificación y regresión [37]. Los valores para la búsqueda de hiperparámetros se utilizó los parámetros: estimator=rf_model, param_distributions=param_grid, scoring='f1_weighted', cv=5, n_iter=50, verbose=1, random_state=42, n_jobs=-. La fórmula del presente modelo se muestra en la ecuación (5):

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_N(x)\} \quad (5)$$

Donde:

- mode: Devuelve la clase más votada entre los árboles
- h_t : Predicción del t-ésimo árbol para la instancia x.
- Naive Bayes
- El modelo de clasificación se basa en el teorema de Naive Bayes. Esto debido a su simplicidad y eficiencia en el procesamiento de grandes volúmenes de datos de texto plano, es muy adecuado para tareas de análisis de sentimientos [38]. En este modelo no aplica búsqueda de hiperparámetros directamente, esto debido a que no tiene muchos

hiperparámetros ajustables que se pueda hacer. La fórmula del presente modelo se muestra en la ecuación (6):

$$P(C_k | x) = \frac{P(C_k) \cdot P(x|C_k)}{P(x)} \quad (6)$$

Donde:

- $P(C_k | x)$: Probabilidad posterior de la clase C_k dada la instancia x .
- $P(C_k)$: Probabilidad a priori de la clase C_k .
- $P(x | C_k)$: Probabilidad condicional de las características x dado que la clase es C_k .
- $P(x)$: Probabilidad total de las características x (normalizador).

2.4.4. Regresión Logística

Es un modelo de tipo supervisado por el cuál se usa para resolver problemas de clasificación basados en la probabilidad de pertenencia a una clase determinada. Está basada en una transformación lógica que convierte combinaciones lineales de variables en probabilidades [39]. La fórmula del presente modelo se muestra en la ecuación (7):

$$P(y = 1 | x) = \frac{1}{1 + e^{-(\omega^T x + b)}} \quad (7)$$

Donde:

- $x = (x_1, x_2, \dots, x_d)$: Vector de características.
- $\omega = (\omega_1, \omega_2, \dots, \omega_d)$: Vector de pesos asociados a las características.
- b : Sesgo o término independiente.
- $\omega^T x = \sum_{i=1}^d \omega_i x_i$: Combinación lineal de características.

2.4.5. K-Nearest Neighbors

Es un modelo de aprendizaje basado en instancias, que se utiliza para la clasificación y regresión. Funciona detectando los “k” vecinos más cercanos de un dataset nuevo basándose en una medida de similitud, como la distancia euclidiana, finalmente asigna una etiqueta de clase en función de las etiquetas mayoritarias de esos vecinos [40]. Los valores para la búsqueda de hiperparámetros se utilizó los parámetros: estimator=knn_model, param_grid=param_grid, scoring='f1_weighted', cv=5, verbose=1, n_jobs=-1. La fórmula del presente modelo se muestra en la ecuación (8):

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (8)$$

Donde:

- $d(x, y)$ es la distancia euclidiana entre dos puntos x y y
- x_i y y_i son los valores de las características i -ésimas de los puntos x y y , respectivamente
- n es el número de características o dimensiones.

2.5. Construcción del Metamodelo

Se construyó un metamodelo basado en boosting mediante stacking ensemble, que es una técnica que utiliza múltiples modelos de forma individual para poder generar predicciones, de las cuáles luego son combinadas por un modelo de nivel superior o meta-modelo [41], [42], [43]. El ensamble boosting es aplicado ya cuando utilizamos modelos como XGBoost y LightGBM. Para evaluar la eficacia de los distintos modelos descritos anteriormente y el meta-modelo de clasificación basado en ensamble boosting se proponen las métricas como la precisión, accuracy, recall, F_1 -score y el AUC basada en la matriz de confusión (Tabla 1). Todas las métricas propuestas anteriormente se calculan utilizando los componentes de la matriz de confusión, así como se exponer en la figura 2 titulada “Arquitectura del Metamodelo de Ensamble para la Clasificación de la Percepción del Servicio”.

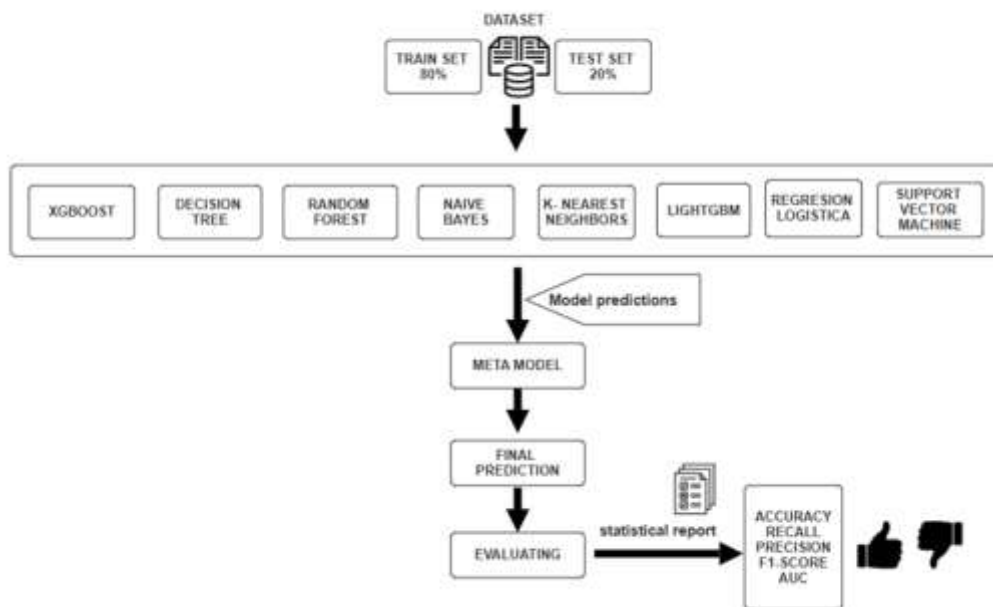


Figura 2. Arquitectura del Metamodelo de Ensamble para la Clasificación de la Percepción del Servicio

2.5.1. Matriz de confusión:

Positive	TP	FN
Negative	FP	TN
	Positive	Negative

Tabla 1. Representación de la Matriz de Confusión para Evaluación de Modelos de Clasificación

La exactitud se puede medir utilizando la ecuación (1)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

La precisión mostrada en la ecuación (2), es definida como la proporción de predicciones positivas correctas (TP), sobre el total de predicciones positivas realizadas (TP+FP).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

El recall mostrada en la ecuación (3) es definida por la proporción de verdaderos positivos (TP) sobre el total de instancias que realmente pertenecen a la clase positiva (TP+FN).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

El F1-score mostrada en la ecuación (4) es la media armónica de la precisión y el recall, lo que significa que da más peso a los valores bajos de estas métricas.

$$F_1-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (4)$$

Y por último el AUC mostrada en la ecuación (5) y (6) es una métrica mediante la cual mide la capacidad de un modelo de clasificación para distinguir entre clases. Se utiliza un conjunto de la curva ROC, la cuál muestra un gráfico de relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

$$TPR = \frac{TP}{TP+FN} \quad (5)$$

$$FPR = \frac{FP}{FP+TN} \quad (6)$$

3. RESULTADOS Y DISCUSIÓN

El análisis utilizó Python como lenguaje de programación principal para generar el meta-modelo y realizar el análisis de interpretación adecuando, añadido a este se implementó la librería de Pyspark (versión 3.5.5) para la carga y creación de dataframes. Esto debido a que Pyspark tiene un mejor manejo de datos a gran escala y permite la fácil manipulación de transformación y limpieza de datos de manera óptima [44]. En la investigación, se identificó los patrones de percepción que tiene la población peruana sobre la calidad del servicio de internet.

3.1. Análisis descriptivo de la percepción del usuario

El gráfico de cajas mostrada en la Figura 3 titulada “Arquitectura del Metamodelo de Ensamble para la Clasificación de la Percepción del Servicio”, nos muestra la distribución de la percepción del servicio de internet en las regiones de Costa, Sierra y Selva previo al balanceo y transformación de datos recolectados de la base de datos. El parámetro P510A, el cual representa el nivel de satisfacción del servicio de internet (1 = Muy insatisfecho/a, 4 = Muy satisfecho/a), presenta diferencias significativas entre regiones. En la región de la costa, la percepción del servicio de internet se centra principalmente en la respuesta 3 (Satisfecho/a), con baja variabilidad y algunos valores irregulares en los extremos (1 y 4), lo cual demuestra una opinión más igualitaria entre los usuarios encuestados. Por lo contrario, en las regiones de la Sierra y la Selva, la distribución es más dispersa, tomando todo el rango de valores, con un valor de mediana en 3 pero con un mayor de respuestas en 1 (Muy insatisfecho/a) y 2 (Insatisfecho/a), lo cual nos da a entender una percepción más heterógena del servicio de estas zonas del país. Tales resultados preliminares sugieren que la satisfacción con los servicios de la red varía según las regiones, siendo así las zonas costeras más estables mientras que las zonas montañosas y selváticas son más diversas, lo que identifica la importancia de analizar los factores que influyen en estas diferencias con más detalles antes de aplicar técnicas de equilibrio o normalización.

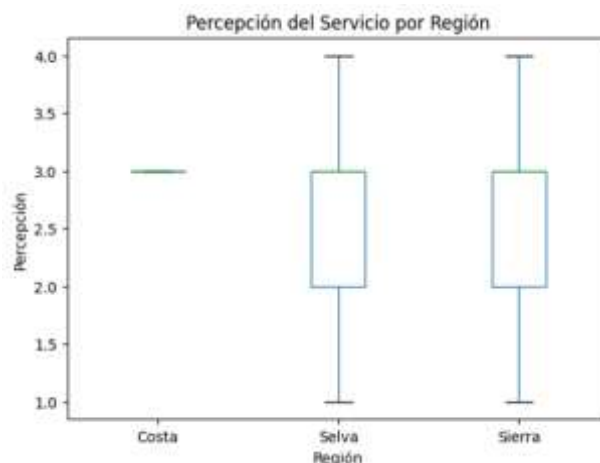


Figura 3. Distribución de la Percepción del Servicio de Internet por Región

Continuamente el gráfico de líneas mostrada en la figura 4 titulada “Distribución de la Percepción del Servicio de Internet por Estrato Socioeconómico”, detalla la distribución de la percepción del servicio de internet según el estrato económico, dando un realce

que en los estratos D (28.6%) y E (26.0%) presentando la mayor proporción de respuestas, lo cuál sugiere que la percepción de los usuarios es más relevante en estos grupos, esto posiblemente debido a problemas de conectividad o calidad del servicio. Por el contrario, los estratos más altos, como lo son el A (6.9%) y B (14.9%), tienen una menor representación, lo que podría indicar un acceso más estable y menos variabilidad en la percepción del servicio. Podemos observar que existe un incremento alto en el Estrato C (23.5%), lo que sugiere también representa una parte importante de la muestra extraída. Esta tendencia general sugiere que los niveles socioeconómicos más bajos participan más en la evaluación de servicios, lo que puede indicar una brecha digital dónde las variaciones en la calidad del acceso a Internet tienen un efecto más notorio en estos sectores.

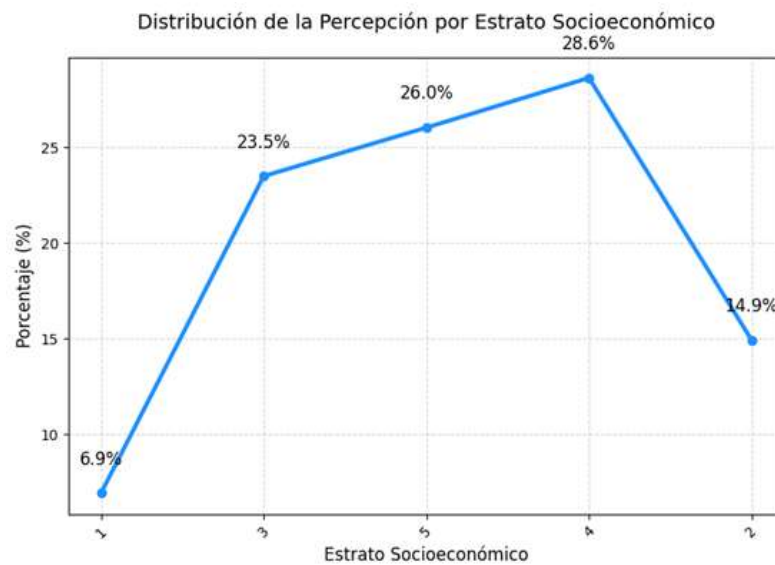


Figura 4. Distribución de la Percepción del Servicio de Internet por Estrato Socioeconómico

3.2. Resultado de la clasificación de patrones

Los resultados de las métricas del entrenamiento de los diferentes modelos en la figura 5 titulada “Comparación de la Precisión de los Modelos de Clasificación Antes y Después del Ajuste”, muestran un incremento y mejora respecto al antes del balanceo de datos con SMOTE y optimización de hiperparámetros con RandomizedSearch. Esto significa que han mejorado significativamente al aplicar estas técnicas al momento de entrenar dichos modelos. Antes de aplicar dichos ajustes, los modelos mostraban una precisión muy cercana al 45%, lo cuál muestra la clara evidencia de sesgo y ruido en las predicciones debido al desbalance total de clases. Sin embargo, después de la implementación de estas técnicas, se puede observar un incremento significativo en la capacidad de clasificación, destacando los modelos de XGBoost, LightGBM y el Metamodelo, que alcanzaron niveles superiores al 67%, consolidándose, así como los modelos más robustos y precisos en el presente estudio. A comparación de modelos como el de Naive Bayes o SVM, presentaron ciertas mejoras, sin embargo, su rendimiento sigue siendo inferior en comparación con los modelos de ensamble. Dichos resultados, destacan la importancia de la correcta selección de parámetros y el

adecuado preprocesamiento, esto para maximizar el desempeño de algoritmos de aprendizaje automático. En definitiva, el empleo de modelos de boosting y enfoques en ensamble tan solo no mejora la capacidad de generalización, sino que también nos permite abordar con mayor precisión problemas de clasificación en escenarios con distribución de datos desiguales, lo cual los convierte en herramientas clave para el análisis de la percepción del servicio y la toma de decisión basada en datos.

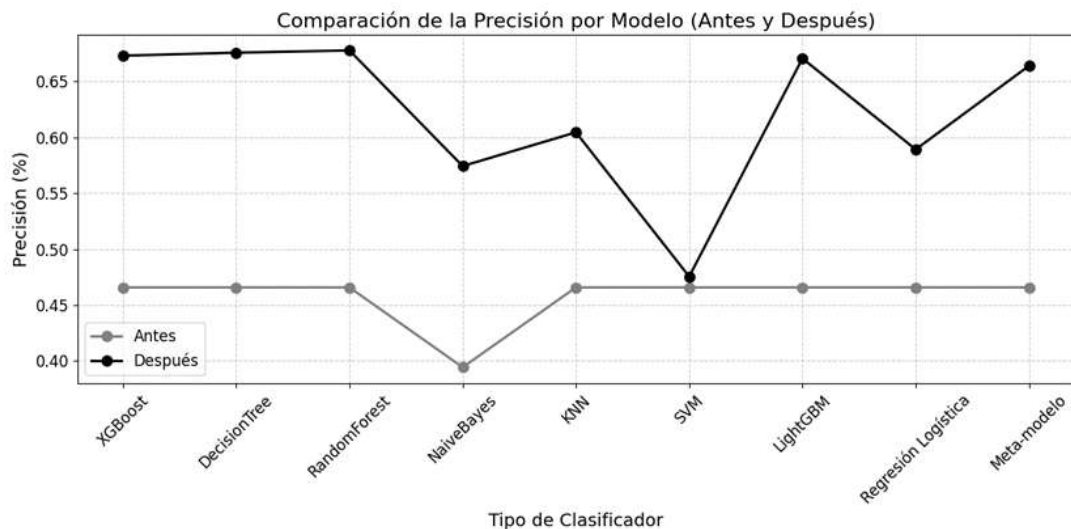


Figura 5. Comparación de la Precisión de los Modelos de Clasificación Antes y Después del Ajuste.

Comparando ahora la métrica del Área Bajo la Curva (AUC) del antes y después, como se muestra en la figura 6 titulada “Comparación del AUC de los Modelos de Clasificación Antes y Después del Ajuste”, podemos ver una mejora significativa similar a la métrica anterior en estos modelos de clasificación. Antes de aplicar las técnicas ya mencionadas anteriormente, los modelos mostraban un rendimiento variable, como lo es XGBoost, Decision Tree y KNN, reflejaban una baja capacidad de diferenciación entre clases. Tras la optimización, se muestra un incremento notable en la mayoría de los modelos, con XGBoost, LightGBM y el Metamodelo destacando de esta manera como las opciones más efectivas, alcanzando valores de AUC superiores a 0.85, lo que indica una mayor capacidad de predicción y robustez en la clasificación. Dichos resultados dan a evidenciar la importancia de aplicar técnicas de preprocesamiento y optimización, al igual que el anterior, esto para poder permitir que los modelos aprendan de manera más equilibrada y mejoren su capacidad de generalización. En conclusión, la combinación de balanceo de datos y ajuste de hiperparámetros he sido clave para lograr una clasificación más precisa y confiable, reduciendo el sesgo en los datos y mejorando la estabilidad de los modelos predictivos.

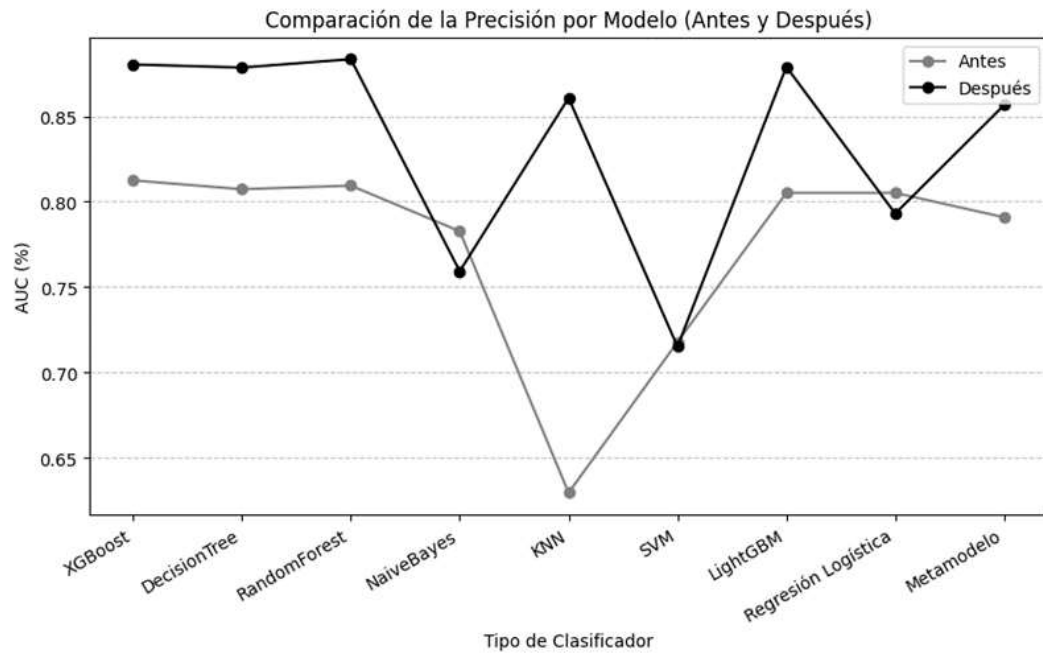


Figura 6. Comparación del AUC de los Modelos de Clasificación Antes y Después del Ajuste.

3.3. Resultados de la clasificación de la percepción del servicio de Internet (comparación de métricas)

Las técnicas aplicadas en esta investigación tal cual se muestra en la tabla 2 titulada “Comparación de las métricas de rendimiento de los modelos de clasificación”, para poder mejorar y optimizar el rendimiento de los modelos propuestos para el presente estudio, se aplicaron de manera exitosa. En particular, XGBoost, LightGBM y Random Forest son los que destacan, alcanzando un AUC de 0.88 y un equilibrio sólido entre precisión, recall y F1-Score, lo que indica una alta capacidad para diferenciar correctamente las clases. El metamodelo también demostró tener un desempeño competitivo con un AUC de 0.86, lo cual reafirma que la combinación de múltiples modelos contribuye a mejorar la estabilidad de las predicciones. Modelos como KNN y Regresión Logística mostraron un desempeño aceptable, pero con menos capacidad de predicción a comparación de los algoritmos basados en boosting. Dichos hallazgos resaltan la importancia de técnicas avanzadas de preprocesamiento y optimización en la construcción de modelos más precisos y confiables, reafirmando que los modelos de boosting y enfoques de ensemble son las estrategias más efectivas para este tipo de clasificación.

Modelo	Accuracy	Recall	Precisión	F1-Score	AUC
XGBoost	0.64	0.64	0.67	0.64	0.88
Random Forest	0.65	0.65	0.68	0.65	0.88

Decision Tree	0.65	0.65	0.68	0.65	0.88
Naive Bayes	0.46	0.46	0.57	0.43	0.76
KNN	0.59	0.59	0.6	0.6	0.86
SVM	0.47	0.47	0.48	0.47	0.72
LightGBM	0.64	0.64	0.67	0.64	0.88
Regresión Logística	0.51	0.51	0.51	0.5	0.79
Metamodelo	0.62	0.62	0.66	0.62	0.86

Tabla 2. Comparación de las métricas de rendimiento de los modelos de clasificación

El presente estudio evaluó como las distintas técnicas de balanceo de datos y el ajuste de hiperparámetros pudo mejorar el rendimiento de cada modelo entrenado. De igual manera sucedió con el metamodelo que combinó múltiples modelos para de esta manera potenciar el resultado de las métricas propuestas para evaluar este estudio y que tan efectivo resultaría ser para la presente investigación, dando como resultado un valor superior como lo es al 0.85 de AUC. De entre estos algoritmos destacan el XGBoost, LightGBM, Random Forest y Metamodelo de Stacking, siendo estos mismos que lograron las mejores precisiones en el entrenamiento. Todos lo ya antes mencionado, destaca la importancia del preprocesamiento de los datos, asimismo, el uso de un metamodelo de ensamble se consolidó como una estrategia eficaz para potenciar la clasificación y obtener resultados más robustos.

3.4. Comparación del modelo Ensamble Boosting y Modelos Individuales

Los modelos de ensamble boosting, como XGBoost y LightGBM, han demostrado ser muy efectivos en tareas de clasificación, superando en muchos casos a los modelos tradicionales. En estudios previos, Haddadi y Hamidi [45] aplicaron Stacking para predecir el Customer Lifetime Value, logrando mejorar la precisión en un 8% respecto a los modelos individuales. De manera similar, Xie et al. [46] usaron Stacking en el análisis de estabilidad en sistemas eléctricos y encontraron que este enfoque aumentó el AUC en un 12% en comparación con Random Forest, lo que refuerza la idea de que combinar modelos puede generar predicciones más solidas. En la predicción de lluvias, Meti et al. [47] demostraron que Stacking obtuvo un AUC de 0.91, mientras que Random Forest y Decision Tree quedaron en 0.88 y 0.85, respectivamente, resultados que coinciden con lo observado en este estudio, donde XGBoost y LightGBM lograron un AUC de 0.88 y precisiones del 64% y 65%, respectivamente. Sin embargo, el metamodelo de Stacking alcanzó un AUC de 0.86 y una precisión del 62%, quedando por debajo de los mejores modelos individuales. Esto sugiere que la configuración del meta-modelo podría mejorarse, lo que también han señalado otros investigadores. Muslim et al. [48], por ejemplo, optimizaron LightGBM con algoritmos de enjambre, elevando su precisión hasta un 95% con ACO y 94.7% con BCO, lo que sugiere que afinar los hiperparámetros puede marcar una gran diferencia. Además, Moutib et al. [49] compararon diferentes modelos para la clasificación de señales biomédicas y hallaron que Random Forest superó a ANN, SVM y regresión logística con una precisión del 92%, lo cual se alinea con nuestros resultados, donde Random Forest también mostró métricas competitivas

(AUC = 0.88, precisión = 65%). Finalmentel, Párvin y Saleena [50] trabajaron con un modelo híbrido de Stacking en la clasificación de sepsis y lograron un AUC de 0.94, lo que sugiere que una selección cuidadosa de modelos y un buen meta-modelo pueden hacer que Stacking supere a los modelos base. En general, estos antecedentes refuerzan que el aprendizaje en ensamble puede mejorar la clasificación, pero también dejan claro que la clase está en ajustar bien los modelos base y el meta-modelo. Para futuros trabajos, podría ser útil probar otras configuraciones en Stacking, explorar diferentes modelos base o aplicar técnicas de optimización de hiperparámetros para mejorar el rendimiento general.

En cuanto a los resultados dados en la métrica de precisión (accuracy), especialmente tras el ajuste del pipeline, se alinean con hallazgos en investigaciones recientes que de igual manera han reportado mejoras significativas. Por ejemplo, en el ámbito biomédico, Asadi et al. [51] mejoraron la precisión de clasificación de tareas de imaginación motora en un 59.81% a 85.19% mediante la integración del algoritmo Common Spatial Pattern con modelos clasificadores SVM y LDA. De forma similar, Hellara et al. [52] alcanzaron una precisión del 88.5% al fusionar señales sEMG y FMG utilizando un algoritmo de selección de características basado en enjambre caótico. En el área de visión por computadora, Wojcuik et al. [53] lograron una mejora de precisión de hasta 6 puntos porcentuales al aplicar técnicas de optimización de hiperparámetros en modelos CNN ajustados. Por su parte, Martins et al. [54] reportaron una precisión de hasta 99.54% al aplicar técnicas de segmentación semántica en archivos ejecutables PE para la clasificación de malware. Finalmente, Xing [55] alcanzó una precisión del 95.38% en la clasificación de datos financieros mediante un modelo de regresión lineal múltiple y selección de variables con valor de Shapley. En comparación, los modelos evaluados en esta investigación, como LightGBM, XGBoost y el metamodelo de ensamble, presentan precisiones que, si bien son más modestas (por encima del 60% tras el ajuste), representan una mejora considerable respecto a la precisión base (~45 %) y validan el impacto positivo de una adecuada ingeniería de características y configuración de hiperparámetros en problemas de clasificación multiclase con datos sociales.

4. CONCLUSIÓN

En este estudio, se propuso un enfoque basado en modelos de clasificación con técnicas de ensamble para poder mejorar la precisión de clasificación de la percepción del servicio de internet. Los resultados demostraron que los algoritmos de boosting (XGBoost y LightGBM) y el metamodelo de Sacking lograron las mejores métricas, alcanzando valores de AUC superiores al 0.85. De igual manera, la combinación de múltiples clasificadores en un metamodelo de ensamble nos permitió obtener una mayor estabilidad y precisión en las predicciones. Asimismo, el análisis de los datos reflejó que una gran parte de los usuarios reporta una percepción negativa del servicio de internet, especialmente en estratos socioeconómicos más bajos, donde la calidad del servicio es más variable y menos confiable, por el contrario, los usuarios con estratos más altos perciben una mejor percepción, demostrando una clara brecha digital en la calidad del servicio de red. En general, el enfoque de ensamble propuesto demostró ser una estrategia efectiva para evaluar la calidad del servicio de internet desde la perspectiva

del usuario, proporcionando así una información clave para las futuras mejoras en la infraestructura de telecomunicaciones.

5. REFERENCIAS

- [1] A. D. Herlambang and A. Rachmadi, "Student's Perception of Technology-Rich Classrooms Usage to Support Conceptual and Procedural Knowledge Delivery in Higher Education Computer Science Course," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1500–1509. doi: 10.1016/j.procs.2024.03.151.
- [2] J. Chen, "Impacts of Internet literacy and Internet contact on the communication effect of university students' ideological and political education in China," *Acta Psychol (Amst)*, vol. 247, Jul. 2024, doi: 10.1016/j.actpsy.2024.104321.
- [3] A. Petrovčič, A. Quan-Haase, B. C. Reisdorf, Š. Žadník, S. Hvalič-Touzery, and J. Laznik, "Categorical and resource inequalities in self-reliant internet use and use-by-proxy among older adults in Slovenia during the COVID-19 pandemic," *Technol Soc*, vol. 79, Dec. 2024, doi: 10.1016/j.techsoc.2024.102735.
- [4] A. R. T. Hidayat, K. Onitsuka, C. P. M. Sianipar, M. Basu, and S. Hoshino, "To migrate or not to migrate: Internet use and migration intention among rural youth in developing countries (case of Malang, Indonesia)," *Digital Geography and Society*, vol. 4, Jan. 2023, doi: 10.1016/j.diggeo.2023.100052.
- [5] R. S. Abd-Ali, S. A. Radhi, and Z. I. Rasool, "A survey: The role of the internet of things in the development of education," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 215–221, 2020, doi: 10.11591/ijeecs.v19.i1.pp215-221.
- [6] M. Martínez-Domínguez and J. Mora-Rivera, "Internet adoption and usage patterns in rural Mexico," *Technol Soc*, vol. 60, Feb. 2020, doi: 10.1016/j.techsoc.2019.101226.
- [7] X. Gao and J. Li, "The nexus between internet use and consumption diversity of rural household," *Heliyon*, vol. 10, no. 15, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35433.
- [8] B. Fomba Kamga, D. N. D. Talla Fokam, and T. N. Nchofoung, "Internet access and innovation in developing countries: some empirical evidence," *Transnational Corporations Review*, Sep. 2022, doi: 10.1080/19186444.2022.2082227.
- [9] M. Rabbani, C. A. Bogulski, H. Eswaran, and C. J. Hayes, "Willingness to pay for internet speed and quality," *Telematics and Informatics*, vol. 93, Sep. 2024, doi: 10.1016/j.tele.2024.102173.
- [10] R. D. Alania-Contreras et al., "Evolving attitudes toward online education in Peruvian university students: A quantitative approach," *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e30566.
- [11] M. De la Cruz and A. Mergoni, "Assessing the performance of Peruvian education system from a governance perspective," *Socioecon Plann Sci*, vol. 93, Jun. 2024, doi: 10.1016/j.seps.2024.101891.

- [12] O. Malamud, S. Cueto, J. Cristia, and D. W. Beuermann, "Do children benefit from internet access? Experimental evidence from Peru," *J Dev Econ*, vol. 138, pp. 41–56, May 2019, doi: 10.1016/j.jdeveco.2018.11.005.
- [13] G. N. C. Sisa, J. G. A. Condori, and W. R. O. Aparicio, "Information and communication technologies, interculturality and rural development in the province of la union, Arequipa Peru," *Publicaciones de la Facultad de Educacion y Humanidades del Campus de Melilla*, vol. 50, no. 2, pp. 15–29, 2020, doi: 10.30827/publicaciones.v50i2.13940.
- [14] A. P. Calatayud Mendoza, B. A. Llerena Zea, E. V. Quispe, P. S. Chialchia, and L. R. Flores Quispe, "Impact of Internet Access on Household's Income in the Department of Puno, Peru," *Lecturas de Economia*, no. 99, pp. 175–207, Jul. 2023, doi: 10.17533/udea.le.n99a352419.
- [15] N. F. M. Zariman, N. Humaidi, and M. H. Abd Rashid, "Mobile commerce applications service quality in enhancing customer loyalty intention: mediating role of customer satisfaction," *Journal of Financial Services Marketing*, vol. 28, no. 4, pp. 649–663, Dec. 2023, doi: 10.1057/s41264-022-00190-9.
- [16] M. Rabbani, C. A. Bogulski, H. Eswaran, and C. J. Hayes, "Willingness to pay for internet speed and quality," *Telematics and Informatics*, vol. 93, Sep. 2024, doi: 10.1016/j.tele.2024.102173.
- [17] S. M. Shrestha and A. Shakya, "A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 652–661. doi: 10.1016/j.procs.2022.12.067.
- [18] S. Chowdhury, A. K. Saha, and D. K. Das, "Hydroelectric Power Potentiality Analysis for the Future Aspect of Trends with R2 Score Estimation by XGBoost and Random Forest Regressor Time Series Models," *Procedia Comput Sci*, vol. 252, pp. 450–456, 2025, doi: 10.1016/j.procs.2025.01.004.
- [19] M. S. Rao, S. Nazma, K. N. Chaitanya, and T. Ambica, "Anomaly based intrusion detection using ensemble machine learning and block-chain," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 3, pp. 2754–2762, Sep. 2024, doi: 10.11591/ijai.v13.i3.pp2754-2762.
- [20] P. Filipe, P. Ruivo, and T. Oliveira, "Assessing machine learning adoption at the firm level: The moderating effect of the environmental context," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 1034–1042. doi: 10.1016/j.procs.2023.01.381.
- [21] M. Vaudevan, R. S. Narayanan, S. F. Nakeeb, and Abhishek, "Customer churn analysis using XGBoosted decision trees," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 488–495, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp488-495.
- [22] N. Hafidi, Z. Khoudi, M. Nachaoui, and S. Lyaqini, "Enhanced SMS spam classification using machine learning with optimized hyperparameters," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 1, pp. 356–364, Jan. 2025, doi: 10.11591/ijeecs.v37.i1.pp356-364.

- [23] N. Santoso, W. Wibowo, and H. Himawati, "Integration of synthetic minority oversampling technique for imbalanced class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 102–108, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.
- [24] M. H. R. Kholiq, Wiranto, and S. W. Sihwi, "News classification using light gradient boosted machine algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 206–213, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp206-213.
- [25] K. Dissanayake and M. G. M. Johar, "Two-level boosting classifiers ensemble based on feature selection for heart disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, pp. 381–391, Oct. 2023, doi: 10.11591/ijeecs.v32.i1.pp381-391.
- [26] A. K. Hamoud et al., "A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 2, pp. 1105–1116, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1105-1116.
- [27] Instituto Nacional de Estadística e Informática (INEI), "Portal de Microdatos," <https://proyectos.inei.gob.pe/microdatos/>.
- [28] C. Chaiya, "Empowering climate resilience: A people-centered exploration of Thailand's greenhouse gas emissions trading and sustainable environmental development through climate risk management in community forests," *Heliyon*, vol. 11, no. 2, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41844.
- [29] P. Fernandes and M. C. Coelho, "Data mining techniques for estimating instantaneous NOX emissions from Euro 6 diesel passenger cars," in *Transportation Research Procedia*, Elsevier B.V., 2024, pp. 444–451. doi: 10.1016/j.trpro.2024.02.056.
- [30] R. A. Casonatto, T. D. P. G. Souza, and A. M. Mariano, "Quality and Risk Management in Data Mining: A CRISP-DM Perspective," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 161–168. doi: 10.1016/j.procs.2024.08.257.
- [31] D. Lee and H. Kim, "The effects of network neutrality on the diffusion of new Internet application services," *Telematics and Informatics*, vol. 31, no. 3, pp. 386–396, Aug. 2014, doi: 10.1016/j.tele.2013.10.001.
- [32] M. Czajkowski, W. Zawadzki, G. Bernatek, and M. Sobolewski, "Assessing the substitutability of mobile and fixed internet: The impact of 5G services on consumer valuation and price elasticity," *Telecomm Policy*, Nov. 2024, doi: 10.1016/j.telpol.2024.102869.
- [33] T. V Geetha and S. Sendhilkumar, "Machine Learning; Concepts, Techniques and Applications," 2023.
- [34] J. Eberlein, D. Rodriguez, and R. Harrison, "The effect of data complexity on classifier performance," *Empir Softw Eng*, vol. 30, no. 1, Feb. 2025, doi: 10.1007/s10664-024-10554-5.
- [35] A. M. Alsaffar, M. Nouri-Baygi, and H. M. Zolbanin, "Shielding networks: enhancing intrusion detection with hybrid feature selection and stack ensemble learning," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00994-7.

- [36] M. Alizamir, M. Wang, R. M. A. Ikram, K. O. Ahmed, S. Heddami, and S. Kim, "An efficient computational investigation on accurate daily soil temperature prediction using boosting ensemble methods explanation based on SHAP importance analysis," *Results in Engineering*, vol. 24, Dec. 2024, doi: 10.1016/j.rineng.2024.103220.
- [37] T. B. Mashifane, L. Braby, M. Pikiso, S. Sunnassee–Taukoor, R. S. Rapolaki, and M. N. Ragoasha, "Machine learning algorithm reveals surface deoxygenation in the Agulhas Current due to warming," *Prog Oceanogr*, vol. 231, Feb. 2025, doi: 10.1016/j.pocean.2024.103407.
- [38] E. S. Negara et al., "SENTIMENT ANALYSIS OF PUBLIC OPINION ON PRESIDENTIAL ADVISORY APPOINTMENTS USING NAIVE BAYES CLASSIFICATION," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 8, no. 2, pp. 452–466, Dec. 2024, doi: 10.22437/jiituj.v8i2.35254.
- [39] J. Á. Martín-Baos, R. García-Ródenas, L. Rodríguez-Benitez, and M. Bierlaire, "Scalable Kernel Logistic Regression with Nyström Approximation: Theoretical Analysis and Application to Discrete Choice Modelling," Feb. 2024, doi: 10.1016/j.neucom.2024.128975.
- [40] J. Zhang, Y. Li, F. Shen, Y. He, H. Tan, and Y. He, "Hierarchical text classification with multi-label contrastive learning and KNN," *Neurocomputing*, vol. 577, Apr. 2024, doi: 10.1016/j.neucom.2024.127323.
- [41] A. Daza et al., "Stacking ensemble based hyperparameters to diagnosing of heart disease: Future works," *Results in Engineering*, vol. 21, Mar. 2024, doi: 10.1016/j.rineng.2024.101894.
- [42] A. M. Haddadi and H. Hamidi, "A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm," *Computers in Human Behavior Reports*, vol. 18, May 2025, doi: 10.1016/j.chbr.2025.100616.
- [43] Z. Xie, D. Zhang, X. Han, and W. Hu, "Power system transient stability preventive control optimization method driven by Stacking Ensemble Learning," *Energy Reports*, vol. 9, pp. 757–765, Oct. 2023, doi: 10.1016/j.egyr.2023.05.106.
- [44] V. Sharma, S. Kannan, S. Tanya, and N. Panda, "Detecting Plant Diseases at Scale: A Distributed CNN Approach with PySpark and Hadoop," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1044–1057. doi: 10.1016/j.procs.2024.04.099.
- [45] A. M. Haddadi and H. Hamidi, "A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm," *Computers in Human Behavior Reports*, vol. 18, May 2025, doi: 10.1016/j.chbr.2025.100616.
- [46] Z. Xie, D. Zhang, X. Han, and W. Hu, "Power system transient stability preventive control optimization method driven by Stacking Ensemble Learning," *Energy Reports*, vol. 9, pp. 757–765, Oct. 2023, doi: 10.1016/j.egyr.2023.05.106.
- [47] G. Meti, R. K. G. Krishnegowda, and G. S. Swamy, "Rainfall analysis and prediction using ensemble learning for Karnataka State," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1187–1198, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp1187-1198.

- [48] M. A. Muslim, Y. Dasril, M. Sam'an, and Y. N. Ifriza, "An improved light gradient boosting machine algorithm based on swarm algorithms for predicting loan default of peer-to-peer lending," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 2, pp. 1002–1011, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1002-1011.
- [49] M. Moutaib, M. Fattah, Y. Farhaoui, B. Aghoutane, and M. El Bakkali, "Fetal electrocardiogram prediction using machine learning: a random forest-based approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1076–1083, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1076-1083.
- [50] S. P. A. and S. B., "Designing a hybrid stack ensemble model to enhance sepsis classification using data triangulation approach," *Results in Engineering*, vol. 25, Mar. 2025, doi: 10.1016/j.rineng.2024.103748.
- [51] O. Asadi, M. Hajihosseini, S. Shirzadi, Z. Einalou, and M. Dadgostar, "Improvement of classification accuracy of functional near-infrared spectroscopy signals for hand motion and motor imagery using a common spatial pattern algorithm," *Intelligent Medicine*, 2025, doi: 10.1016/j.imed.2024.05.004.
- [52] H. Hellara, R. Barioul, S. Sahnoun, A. Fakhfakh, and O. Kanoun, "Improving the accuracy of hand sign recognition by chaotic swarm algorithm-based feature selection applied to fused surface electromyography and force myography signals," *Eng Appl Artif Intell*, vol. 154, Aug. 2025, doi: 10.1016/j.engappai.2025.110878.
- [53] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.
- [54] E. Martins, J. B. Higuera, R. Sant'Ana, J. R. B. Higuera, J. A. S. Montalvo, and D. P. Castillo, "Semantic Malware Classification Using Artificial Intelligence Techniques," *CMES - Computer Modeling in Engineering and Sciences*, vol. 142, no. 3, pp. 3031–3067, 2025, doi: 10.32604/cmes.2025.061080.
- [55] Y. Xing, "Financial Data set Classification Method based on Linear Regression Model," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 970–979. doi: 10.1016/j.procs.2024.10.117.

	Gutierrez Beltrán Alejandro Luis Javier    is a Systems Engineer graduated from Peruvian Union University (UPeU) with four years of experience in software development and data engineering. His career started before finishing university at ADRA as a Junior Full-Stack Developer, later joining Uaiki as a Software Developer, where he worked on a web application platform. He then transitioned into data engineering at Bluetab, working as a Data Developer for a banking client, focusing on big data technologies. After that, he joined Indra as a Data Engineer, continuing his work with the same banking client. Currently, he is part of Inetum, where he applies his knowledge using various methodologies, constantly improving his skills and adapting to Scrum team environments, aiming to grow professionally and contribute to the company's success. He can be contacted at email: alejandro.gutierrez@upeu.edu.pe.
	Fernandez Blanco Arnold    A Systems Engineer from the Peruvian Union University (UPeU), he has over five years of experience specializing in data engineering. He began his career as a Systems Assistant in the DIGETI department at UPeU, where he acquired a solid foundation in IT support and management. He later joined Global Fiber as a Systems Assistant, working as a developer and providing technical support. His transition to data engineering took place at Bluetab, where he worked as a Data Developer for a banking client, specializing in Big Data technologies. During this time, he applied his technical skills to design innovative solutions and contribute to the success of large-scale projects. He currently works at Indra as a Data Engineer, where he continues to strengthen his technical and soft skills. His focus is on collaborating with multidisciplinary teams to achieve business objectives efficiently and strategically. He can be contacted at email: arnoldferb@gmail.com
	Prof. Dr. Soria Quijaite Juan Jesús     PhD in Systems Engineering, Master in University Teaching and Educational Management, master's in applied mathematics and specialist in Statistics for Research. Statistical consultant at the Ministry of Production Cite Agroindustrial, research professor at the Peruvian Union University, member of the global network of researchers AUTHOR AID, member of the Network of Teachers of Latin America and the Caribbean, member of the National System of Science, Technology and Technological Innovation RENACYT (CONCYTEC) at Level VI, thesis advisor with the approaches of the scientific method, good practices of PMBOK version 7 and systems thinking in different specialties of undergraduate and graduate universities. He can be contacted at email: jesussoria@upeu.edu.pe



Prof. MSc Nemias Saboya     (Member, IEEE) received the B.S. degree in systems engineering from Universidad Peruana Unión (UPeU), Peru, and the M.S. degree in computer and systems engineering, with a specialization in information technology management, USMP, Peru. He is currently pursuing the Ph.D. degree in systems engineering with UPeU. His main research interests include data governance, statistical machine learning, IT, and data science, member of the National System of Science, Technology and Technological Innovation RENACYT (CONCYTEC). He can be contacted at email: saboya@upeu.edu.pe.

EVIDENCIA DE ACEPTACION DEL ARTÍCULO CIENTÍFICO

[IJEECS] Accepted!! "Machine learning with boosting on the perception of the Internet service in Peru"

Dear Prof/Dr/Mr/Mrs: Mr Nemias Saboya,

It is my great pleasure to inform you that your paper entitled "Machine learning with boosting on the perception of the Internet service in Peru" is initially ACCEPTED and will be published on the Indonesian Journal of Electrical Engineering and Computer Science later, after all final documents have been completed and reached us. [Congratulations!](#)

Please submit your below documents within 30 days to ijeecs.iaes@gmail.com

1. Final camera ready paper (in MS Word file format, or LaTeX source files)
2. Similarity report checking (that the result is below 25%, uses software such as iThenticate or Turnitin), and
3. Payment evidence

Pay attention: Avoid author self-citations, as well as this journal and the publisher's self-citations!

When you send in all of the above documents and your final paper follows all of the final instructions, your paper will be judged "fully accepted." It will be the date of your paper's acceptance.

We are looking forward to hearing from you

Thank you

"AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA"

RESOLUCIÓN N° 0292-2025/UPeU-FIA-CF

Lima, Naña, 24 de junio de 2025

VISTO:

El expediente de los (las) bachilleres **Arnold Fernández Blanco** identificado(a) con código universitario N° 201410191 y **Alejandro Luis Javier Gutiérrez Beltrán** identificado(a) con código universitario N° 201810196, 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 N° 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 sustentación de la tesis en formato artículo;

Que el Comité Dictaminador ha emitido su dictamen aprobando el informe de tesis titulado "Machine learning con boosting sobre la percepción del servicio de Internet en Perú", presentado por los (las) bachilleres **Arnold Fernández Blanco** y **Alejandro Luis Javier Gutiérrez Beltrán**, reuniendo de esta manera las condiciones previas para la declaración de expedito para la programación de la sustentación;

Estando a lo acordado en la sesión del Consejo de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, celebrada el 24 de junio de 2025, y en aplicación del Estatuto y el Reglamento General de investigación de la Universidad;

SE RESUELVE:

1. Declarar expedito a los (las) bachilleres **Arnold Fernández Blanco** y **Alejandro Luis Javier Gutiérrez Beltrán**, para que sustenten la tesis en formato artículo titulada "Machine learning con boosting sobre la percepción del servicio de Internet en Perú", conducente a la obtención del título profesional de Ingeniero de Sistemas, el 08 de julio a las 09:00 horas, en la modalidad Virtual u online sincrónica.
2. Designar el Jurado de Sustentación, encargado de gestionar la sustentación respectiva, el mismo que queda constituido por los siguientes miembros:

Presidente: Mg. Immer Elías Cuellar Rodríguez
Secretario: Mg. Lenin Henry Centurion Julca
Vocal 1: Mg. Nemius Saboya Ríos
Vocal 2: Mg. Ferdinand Edgardo Pineda Anco
Asesor: Dr. Juan Jesús Soria Quijaite

Regístrese, comuníquese y archívese.




Dra. Enka Inés Acuña Salinas
DECANA

OT:
Informado
Jurado 041
Secretaría General
Archivos




Ph.D. Silvia Pilco Quesada
SECRETARIA ACADÉMICA