



UNIVERSIDAD PERUANA
CAYETANO HEREDIA
ESCUELA DE POSGRADO VÍCTOR ALZAMORA CASTRO

**PREDICCIÓN DE LA VINCULACIÓN
LABORAL EN INVESTIGADORES
DEL ÁREA DE CIENCIAS MÉDICAS Y
DE SALUD DEL REGISTRO
NACIONAL DE CIENCIA Y
TECNOLOGÍA DE PERÚ:
APLICACIÓN DE MODELOS DE
MACHINE LEARNING**

**TESIS PARA OPTAR EL GRADO DE
MAESTRO EN POLÍTICAS Y GESTIÓN
DE LA CIENCIA, TECNOLOGÍA E
INNOVACIÓN**

**MARCOS AUGUSTO, ESPINOLA
SÁNCHEZ**

**LIMA – PERÚ
2025**

ASESOR

MG. YAHIR, DELZO LAZO

JURADO DE TESIS

MG. ASTRID HASSEL, CRIALES JOHNSON

PRESIDENTE

MG. RAQUEL MERCEDES, SOTOMAYOR PARIAN

VOCAL

MG. DAVID ERNESTO, LUJAN TANTARICO

SECRETARIO

DECLARACIÓN DE AUTOR			
FECHA	25	AGOSTO	2025
APELLIDOS Y NOMBRES DEL EGRESADO	MARCOS AUGUSTO ESPINOLA SANCHEZ		
PROGRAMA DE POSGRADO	MAESTRÍA EN POLÍTICAS Y GESTIÓN DE LA CIENCIA, TECNOLOGÍA E INNOVACIÓN		
AÑO DE INICIO DE LOS ESTUDIOS	2022		
TITULO DEL TRABAJO DE INVESTIGACIÓN DE GRADO	"PREDICCIÓN DE LA VINCULACIÓN LABORAL EN INVESTIGADORES DEL ÁREA DE CIENCIAS MÉDICAS Y DE SALUD DEL REGISTRO NACIONAL DE CIENCIA Y TECNOLOGÍA DE PERÚ: APLICACIÓN DE MODELOS DE MACHINE LEARNING"		
MODALIDAD DE TRABAJO DE GRADO	TESIS		
Declaración del Autor El presente Trabajo de Grado es original y no es el resultado de un trabajo en colaboración con otros, excepto cuando así está citado explícitamente en el texto. No ha sido ni enviado ni sometido a evaluación para la obtención de otro grado o diploma que no sea el presente.			
Teléfono de contacto (fijo / móvil)	987654324		
E-mail	Viky.panduro@hotmail.com		



Firma del Egresado
DNI 43525806

DEDICATORIA.

A mi familia, por ser el pilar fundamental en mi camino,
por su apoyo incondicional, comprensión y motivación constante.

A mi padre, Augusto Espinola Flores, cuya calma ante la adversidad se resumía
en
una frase que aún resuena en mí: “¿Qué es lo peor que puede pasar?”.
De él aprendí que, incluso cuando las dificultades ya han ocurrido,
lo esencial son las decisiones que tomamos y
la actitud con la que elegimos continuar.

A mi madre, Estela Sánchez de Espinola, quien, con su ejemplo cotidiano,
me enseñó que alcanzar lo que importa demanda una voluntad firme.

A quien supo acompañarme con paciencia, cercanía y aliento sincero, y que,
incluso en días difíciles, me recordó con gestos simples y constantes que también
existen otras formas de habitar el tiempo: más ordenadas, más conscientes y
plenas.

A cada uno de ellos, gracias por mostrarme que no basta con ser resiliente,
sino que también es posible transformarse, crecer y construir
aquello que realmente importa a partir de cada desafío.

A quienes integran la comunidad científica y de innovación del Perú, que con
compromiso y perseverancia impulsan la generación de conocimiento y
desarrollo, incluso en contextos marcados por limitaciones estructurales e
institucionales.

AGRADECIMIENTOS.

A mi familia y amigos, por su respaldo incondicional
y aliento constante durante cada etapa.

A mi asesor, Yahir Delzo Lazo, por su guía comprometida,
sus orientaciones oportunas y por las enseñanzas
que han enriquecido tanto este trabajo cómo mi formación.

Expreso mi gratitud a mis docentes y compañeros de la Maestría en Políticas y
Gestión de la Ciencia, Tecnología e Innovación, por compartir
con generosidad sus conocimientos, experiencias,
y por contribuir de manera significativa.

FUENTES DE FINANCIAMIENTO.

Tesis Autofinanciada

ÍNDICE GENERAL

CONTENIDO	PÁGINA
PORTADA	
DEDICATORIA	
AGRADECIMIENTOS	
RESUMEN	
ABSTRACT	
I. INTRODUCCIÓN	1
1.1. Antecedentes	1
1.2. Planteamiento del problema	3
1.3. Justificación	5
1.4. Objetivos	7
1.4.1. Objetivo general	7
1.4.2. Objetivos específicos	7
II. METODOLOGÍA	8
2.1. Enfoque	8
2.2. Tipo y diseño	8
2.3. Población y muestra	8
2.3.1. Población	8
2.3.2. Criterios de selección	8
2.3.3. Muestra y tipo de muestreo	8
2.4. Variables	9
2.4.1. Variables independientes (predictoras)	9
2.4.2. Variable dependiente (desenlace)	10
2.5. Procedimientos	10

CONTENIDO	PÁGINA
2.6. Plan de análisis	13
2.6.1. Análisis descriptivo	13
2.6.2. Construcción de modelos predictivos	13
2.7. Consideraciones éticas	14
 III. RESULTADOS	 16
3.1. Descripción de la muestra	16
3.2. Descripción de la producción científica y académica de investigadores RENACYT según vinculación laboral	19
3.3. Modelos predictivos con censura temporal (modelo principal)	21
3.3.1. Análisis de curvas ROC	22
3.3.2. Métricas de rendimiento del modelo Gradient Boosting y comparación con otros algoritmos	23
3.3.3. Feature importance del modelo con mayor rendimiento	25
3.3.4. SHAP values del modelo con mayor rendimiento	26
3.4. Consistencia del modelo de ML con mejor desempeño	30
3.5. Análisis exploratorio con variables extendidas no censuradas	31
3.5.1. Análisis de curvas ROC de modelos	31
3.5.2. Métricas de rendimiento del modelo	33
 IV. DISCUSIÓN	 35
 V. CONCLUSIONES Y RECOMENDACIONES	 41
5.1. Conclusiones	41
5.2. Recomendaciones	41
 VI. REFERENCIAS BIBLIOGRÁFICAS	 43

ÍNDICE DE TABLAS Y FIGURAS

TABLAS

- **Tabla 1.** Caracterización profesional, académica y productiva de investigadores RENACYT del área de Ciencias Médicas y de la Salud afiliados a instituciones académicas, con o sin doble vinculación al sector público o privado 17
- **Tabla 2.** Comparación de métricas de rendimiento de modelos para la predicción de vinculación laboral utilizando variables censuradas temporalmente 24
- **Tabla 3.** Comparación de métricas de rendimiento de modelos para la predicción de vinculación laboral utilizando variables censuradas temporalmente y no censuradas 34

FIGURAS

- **Figura 1.** Flujo de procesamiento de datos y técnicas utilizadas en el modelado mediante machine learning. 12
- **Figura 2.** Distribución de la producción académica y científica de investigadores RENACYT con vinculación solamente académica 19
- **Figura 3.** Distribución de la producción académica y científica de investigadores RENACYT antes de tener vinculación académica y con el sector público o privado 20
- **Figura 4.** Distribución de la producción académica y científica de investigadores RENACYT después de la vinculación académica y con el sector público o privado 21
- **Figura 5.** Curvas ROC comparativas de modelos de machine learning para la predicción de vinculación laboral utilizando variables censuradas temporalmente 23
- **Figura 6.** Importancia relativa de las variables predictoras en el modelo Gradient Boosting (modelo principal, con censura temporal) 26
- **Figura 7.** Gráfico resumen de valores SHAP del modelo Gradient Boosting entrenado con variables censuradas temporalmente 29
- **Figura 8.** Distribución de valores del área bajo la curva ROC en repeticiones de entrenamiento y validación del modelo Gradient Boosting 30
- **Figura 9.** Distribución de valores de F1 score en repeticiones de entrenamiento y validación del modelo Gradient Boosting 31
- **Figura 10.** Curvas ROC comparativas de modelos de machine learning para la predicción de vinculación laboral utilizando variables censuradas temporalmente y no censuradas 32

RESUMEN

Introducción: La vinculación laboral con el sector público o privado es una característica creciente en los ecosistemas científicos de países con recursos limitados. Sin embargo, se desconoce en qué medida ciertos indicadores académicos y profesionales predicen esta condición entre investigadores del área de Ciencias Médicas y de la Salud en Perú. **Objetivo:** Desarrollar y validar modelos de machine learning para predecir la vinculación laboral con el sector público o privado entre los investigadores RENACYT con filiación académica y del área de Ciencias Médicas y de la Salud. **Métodos:** Estudio observacional, transversal, basado en datos de 1,340 investigadores con filiación académica registrados en RENACYT entre 2021 y 2023. Se construyeron modelos predictivos (Gradient Boosting, Random Forest, XGBoost, entre otros) empleando variables profesionales, de producción académica y científica. Se seleccionó el modelo con mejor rendimiento. **Resultados:** El 66.9 % de los investigadores reportó vinculación con entidades del sector público o privado adicional a su filiación académica con la universidad. El modelo principal (Gradient Boosting) alcanzó una AUC de 0.85, exactitud de 0.87, sensibilidad de 0.91 y precisión de 0.90. La mayoría de los investigadores presentó bajo impacto bibliométrico (mediana índice $H=2$), escasa participación en proyectos competitivos (mediana = 0) y limitada dirección de tesis. Las variables del modelo fueron publicaciones indexadas en Scopus, tesis dirigidas y participación en proyectos con financiamiento. Sin embargo, los valores SHAP mostraron efectos no unidireccionales. **Conclusión:** Los perfiles con vinculación académica y con el sector público o privado no se explican únicamente por la productividad científica, sino por la interacción multidimensional entre indicadores profesionales y académicos. El uso de modelos de machine learning y valores SHAP permitió identificar patrones latentes y aportar evidencia útil para la comprensión de trayectorias científicas en contextos con restricciones estructurales. Estos hallazgos pueden informar políticas públicas e institucionales orientadas a fortalecer sectores del ecosistema científico.

Palabras claves: Ciencias de la Salud, Investigadores, Capital Humano, Aprendizaje Automático

ABSTRACT

Introduction: Employment linkage with the public or private sector is an increasingly common feature in scientific ecosystems of resource-constrained countries. However, it remains unclear to what extent certain academic and professional indicators predict this condition among researchers in the field of Medical and Health Sciences in Peru, registered in the National Registry of Science and Technology (RENACYT). **Objective:** To develop and validate machine learning models to predict employment linkage with the public or private sector among RENACYT researchers affiliated with academic institutions in the field of Medical and Health Sciences. **Methods:** A cross-sectional observational study was conducted using data from 1,340 researchers with academic affiliation, registered in RENACYT between 2021 and 2023. Predictive models (Gradient Boosting, Random Forest, XGBoost, among others) were built using professional, academic, and scientific production variables, applying temporal censorship up to the event of interest. The model with the best performance was selected. **Results:** A total of 66.9% of researchers reported employment linkage with public or private sector entities in addition to their academic affiliation. The best-performing model (Gradient Boosting) achieved an AUC of 0.85, accuracy of 0.87, sensitivity of 0.91, and precision of 0.90. Most researchers exhibited low bibliometric impact (median H-index = 2), limited participation in competitive research projects (median = 0), and minimal thesis supervision. Key predictive variables included Scopus-indexed publications, thesis supervision, and participation in funded projects. However, SHAP values revealed non-unidirectional and complex effects. **Conclusion:** Employment linkage with public or private sectors among academic researchers is not solely determined by scientific productivity but results from a multidimensional interaction of academic and professional indicators. The use of machine learning models and SHAP values enabled the identification of latent patterns, offering valuable evidence for understanding research career trajectories in structurally constrained environments. These findings may support the design of public and institutional policies to strengthen the scientific ecosystem.

Keywords: Health Sciences; Researchers; Human Capital; Machine Learning.

INTRODUCCIÓN

1.1. Antecedentes

La incorporación de profesionales con experiencia en investigación académica ha sido reconocida como un componente estratégico para el fortalecimiento de la innovación dentro de organizaciones públicas y privadas. No obstante, esta ventaja competitiva solo se materializa cuando existe una cultura organizacional que estimule activamente la generación de conocimiento científico, condición necesaria para asegurar retornos significativos en innovación (1). A pesar de ello, persisten desafíos en la relación entre la academia y el sector productivo. Mientras que los académicos tienden a priorizar la producción de conocimiento sobre su aplicación comercial, la interacción entre ambos sectores debe orientarse a un equilibrio entre el avance científico y la transferencia tecnológica (2).

Asimismo, la relación entre el perfil de los investigadores y sus condiciones laborales revela dinámicas complejas. Investigaciones recientes han mostrado que los salarios de científicos vinculados al ámbito académico tienden a incrementarse cuando se orientan hacia actividades de desarrollo, pero disminuyen en función del tiempo dedicado exclusivamente a la investigación, lo cual contribuye a una brecha sostenida con respecto al sector industrial (3).

Paralelamente, los modelos de machine learning (ML) han emergido como herramientas eficaces para analizar y predecir resultados en el mercado laboral, particularmente en sectores como la salud y la educación superior (4–7). Su aplicación en contextos académicos y sanitarios ha permitido identificar patrones relacionados con el bienestar ocupacional, incluyendo el burnout y la rotación del personal, aportando evidencia valiosa para optimizar la gestión del talento humano en instituciones clínicas y de investigación (4,8,9).

Estudios han abordado la aplicación de técnicas de ML para analizar la empleabilidad. García-Peñalvo et al. en el contexto del Observatorio de Empleabilidad y Empleo Universitario de España, utilizaron modelos de clasificación supervisada para predecir la empleabilidad a partir de datos institucionales, integrando variables académicas y socioeconómicas (10). De igual modo, el estudio de Clemente C.J. empleó algoritmos de ML, incluyendo Random

Forest y XGBoost, para predecir la empleabilidad de egresados de tecnologías de la información en universidades de América Latina, demostrando una alta capacidad predictiva del modelo (11).

Investigaciones recientes han explorado la aplicabilidad de modelos explicativos y predictivos para anticipar la deserción laboral, así como la adecuación ocupacional y la satisfacción profesional en profesionales sanitarios y docentes (8,12,13). En el campo de la salud mental, Fukui et al. aplicaron modelos de machine learning a datos de recursos humanos de centros comunitarios de salud mental, logrando predecir con efectividad la rotación laboral de los empleados [9]. Mientras que Chen et al. diseñaron un algoritmo de ML basado en datos de smartphones para predecir la probabilidad de trabajo remoto, un aspecto cada vez más relevante en la configuración de entornos laborales híbridos para profesionales de la salud (14).

Gupta et al. desarrollaron modelos de ensamblado para predecir la satisfacción laboral y el comportamiento de empleados en instituciones de Asia, confirmando la utilidad de técnicas como Gradient Boosting para sistemas de recursos humanos basados en datos (15). En relación con la formación médica especializada, Thomas et al. desarrollaron un modelo de red neuronal artificial para predecir la realización personal y la satisfacción profesional de médicos en países en desarrollo, destacando implicancias para la retención de talento (16).

El uso de modelos de predicción laboral mediante ML se ha extendido progresivamente hacia contextos académicos y de salud de alta rotación y demanda profesional (5,7). En el estudio de Hannaford et al. se emplearon modelos de machine learning para predecir la graduación exitosa de estudiantes de enfermería en programas de licenciatura, facilitando decisiones en procesos formativos y de selección (6). Mientras que Habehh y Gohel presentaron una revisión sistemática sobre la aplicación de machine learning en salud, resaltando su potencial para anticipar decisiones laborales en diferentes subespecialidades clínicas (17).

En el contexto nacional, se ha observado un creciente interés en la aplicación de enfoques computacionales para el estudio de la empleabilidad en el Perú. Espinosa-Pinos et al. aplicaron redes neuronales artificiales para predecir la satisfacción laboral en docentes universitarios, demostrando la utilidad de estos modelos en contextos educativos locales (18). De forma complementaria, Charapaqui-Miranda et al. evaluaron la precisión de diversos algoritmos de ML en la predicción de la aptitud laboral durante evaluaciones ocupacionales en trabajadores peruanos; destacando el potencial de la inteligencia artificial para apoyar decisiones en salud ocupacional (19).

Aunque se han realizado estudios exploratorios sobre los currículums de investigadores peruanos registrados en el Directorio Nacional de Investigadores, persiste un vacío importante en el análisis específico de su vinculación laboral. Este aspecto es fundamental para comprender la dinámica del talento humano en ciencia y tecnología en Perú, dado que dicho registro constituye una fuente crítica de información sobre los profesionales del sector investigación en el país (20).

1.2. Planteamiento del problema

El desarrollo, adopción y difusión de innovaciones en el dinámico y complejo sistema del sector de la salud requiere de la integración laboral de investigadores (21). La motivación de las empresas para contratar académicos destacados abarca una variedad de factores que influyen en esta elección, enfatizando especialmente las habilidades tanto en investigación como en desarrollo (22). Por otro lado, la participación desde la academia busca traspasar fronteras como un elemento crucial para el éxito en la transferencia y comercialización de tecnologías universitarias (23).

El personal de salud orientado a investigación científica es fundamental para la formación de centros académicos orientados a la investigación; sin embargo, los estudiantes universitarios e investigadores enfrentan diversos desafíos y barreras con la investigación (24). Estos obstáculos van desde la naturaleza fragmentada de la formación entre la academia y la industria, hasta compromisos temporales entre la industria y la academia (25). En este contexto, resalta la transformación de la educación superior hacia la comercialización de la ciencia, interacción

academia-industria, nuevas prácticas entre científicos, universidades y la industria (26).

La colaboración efectiva entre la academia y la industria se ha vuelto esencial para impulsar el progreso en la investigación y desarrollo (I+D); siendo relevante el perfil que tengan los investigadores (27). La educación en emprendimiento científico se presenta como una respuesta a las crecientes expectativas sobre estudiantes de doctorado y profesores en ciencias e ingenierías. Este modelo, que se basa en factores clave como la preparación tecnológica y las decisiones sobre propiedad intelectual, busca fomentar una participación más activa y consciente de los investigadores en la comercialización de sus descubrimientos (28).

Al resaltar la complejidad de las investigaciones e innovaciones corporativas, se destacan atributos críticos, sugiriendo que los investigadores necesitan habilidades esenciales para el éxito como investigadores corporativos tales como contabilidad y derecho, además de habilidades investigativas (29). Por otro lado, la creciente relevancia de las habilidades analíticas de datos en el entorno laboral actual resalta la necesidad de una preparación más ajustada a las demandas actuales del mercado laboral en los perfiles profesionales (30).

En el Perú, se observa una clara discrepancia en los indicadores de innovación y competitividad global en comparación con países de la OCDE y Latinoamérica. Por ejemplo, el índice de innovación peruano es de 13.8, mientras que en países de la OCDE alcanza los 62.8 y en Latinoamérica llega a 25.6. Además, el puntaje en el índice de competitividad global es de 62, comparado con los 84 de la OCDE y los 71 de Latinoamérica (31).

En el ámbito empresarial peruano, las microempresas representan el 94,2%, seguidas por un 5% de pequeñas empresas, un 0,2% de empresas medianas y un 0,6% de grandes empresas. Se evidencia una clara priorización del trabajo operativo sobre la inversión en investigación y desarrollo. Por ejemplo, la productividad de las empresas medianas es solo el 64,6% de las grandes, y descende aún más en las pequeñas empresas al 27,1%, y en las microempresas al 4,6% (32). En cambio, el desarrollo de investigación e innovación se concentra

mayormente en las grandes empresas (64,8%) y en las pequeñas empresas (29,7%), seguido por las medianas empresas (5.5%) (33).

Para abordar esta brecha, el Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) ha implementado políticas orientadas a fortalecer la investigación y la innovación, promoviendo la articulación entre la academia, la industria y otros actores del ecosistema científico. En este marco, CONCYTEC administra el Registro Nacional Científico, Tecnológico y de Innovación Tecnológica (RENACYT), que constituye el único mecanismo formal de acreditación del Perú para reconocer a una persona como investigador. La incorporación al RENACYT requiere cumplir criterios rigurosos relacionados con la formación académica, la producción científica, la participación en proyectos de investigación y el desarrollo de propiedad intelectual, garantizando así estándares de calidad y mérito científico (34). Además, la SUNEDU también considera mecanismos en donde las universidades cuenten con docentes investigadores RENACYT como parte de su labor de aseguramiento de calidad, lo que evidencia un enfoque en la generación de conocimiento y patentes (35).

En Perú, las acciones de financiamiento parcial o total de actividades de investigación científica, desarrollo tecnológico e innovación se ejecutan mediante mecanismos concursables, bajo estándares técnicos definidos por el CONCYTEC como ente rector del SINACTI (36). La consolidación de un sistema nacional robusto de investigación e innovación requiere un entorno favorable que incluya inversión suficiente en I+D, especialmente del sector privado; instituciones e investigadores de alta calidad; colaboración entre academia e industria; protección efectiva de la propiedad intelectual; y acceso a financiamiento, capital de riesgo y una alta competencia local (37). En el ámbito de la salud, las innovaciones tecnológicas, desde la perspectiva de la economía de la innovación, ofrecen un importante potencial para mejorar la eficiencia del sistema sanitario y reducir costos, lo que refuerza la necesidad de promover la investigación y el desarrollo orientados al avance terapéutico y a la optimización del uso de recursos (21).

Contribuir al análisis de la vinculación laboral de los investigadores en ciencias médicas y de la salud se vuelve relevante en un entorno donde la conexión entre la investigación académica y el desarrollo industrial es limitada (24). Además, fortalecer la transformación académica hacia el emprendimiento y los procesos de transferencia tecnológica se presenta como una estrategia relevante para cerrar esta brecha (23).

1.3. Justificación

El sistema de ciencia, tecnología e innovación del Perú presenta limitaciones estructurales que afectan la sostenibilidad de las trayectorias investigadoras (36). Esta situación se enmarca en una tendencia regional, donde la expansión del sistema de educación superior en América Latina ha estado acompañada por un incremento del empleo académico de carácter parcial, inestable y escasamente remunerado (38). Estas condiciones laborales limitan la consolidación de carreras científicas y fomentan la búsqueda de fuentes complementarias de ingreso entre los investigadores.

En contextos como el europeo, la evidencia empírica ha demostrado que los investigadores que mantienen simultáneamente una vinculación académica y otra en una entidad pública o privada pueden alcanzar un rol estratégico para promover colaboraciones entre sectores. Sin embargo, esta relación no siempre resulta lineal ni uniforme (39). En el caso peruano, el fenómeno de la doble vinculación ha sido escasamente estudiado, y no se dispone de evidencia empírica ni de modelos predictivos que permitan estimar, a partir de características observables, la probabilidad de que un investigador académico mantenga una doble afiliación laboral. Esta carencia limita tanto la comprensión del fenómeno como la posibilidad de diseñar intervenciones informadas desde la política pública o la gestión institucional.

Frente a esta brecha, el presente estudio contribuye al análisis del empleo científico en el país desde un enfoque metodológico y empírico. Para ello, se construyeron y validaron modelos de ML utilizando datos secundarios de investigadores del área de Ciencias Médicas y de la Salud registrados en RENACYT con afiliación académica. El alcance del estudio no se limita a estimar

la frecuencia de la doble vinculación laboral, sino también identificar los patrones asociados a esta condición. Los hallazgos podrán orientar a políticas dirigidas al fortalecimiento de la carrera investigadora, al diseño de incentivos diferenciados según el tipo de vinculación, y a la consolidación de un ecosistema de CTI más articulado, sostenible y acorde con los desafíos del país.

1.4.Objetivos

1.4.1. Objetivo general:

Desarrollar y validar modelos de machine learning para predecir la vinculación laboral con el sector público o privado entre los investigadores RENACYT con filiación académica y del área de Ciencias Médicas y de la Salud.

1.4.2. Objetivos específicos:

- Caracterizar el perfil de los investigadores RENACYT con filiación académica y del área de ciencias médicas y de la salud, según su nivel de categorización RENACYT, grado académico más alto alcanzado, tipo de profesión universitaria, tesis dirigidas por nivel académico, idioma inglés y otros idiomas, índice H, número de proyectos con financiamiento, publicaciones científicas según cuartiles Scopus, así como desarrollos tecnológicos, patentes registradas y desarrollos industriales.
- Diseñar y entrenar modelos de machine learning utilizando distintos enfoques, incluyendo arquitecturas de redes neuronales, para predecir la vinculación laboral al sector público o privado en investigadores con filiación académica.
- Evaluar y validar el rendimiento predictivo de los modelos desarrollados para predecir la vinculación laboral al sector público o privado en investigadores con filiación académica.

II. METODOLOGÍA

2.1. Enfoque: Cuantitativo

2.2. Tipo y diseño:

Se realizó un estudio predictivo usando ML, con un diseño no experimental, observacional, transversal y analítico, basado en el análisis secundario de datos registrados en la plataforma RENACYT entre los años 2021 y 2023. Se emplearon algoritmos de aprendizaje automático para desarrollar y validar modelos que estimaran la probabilidad de vinculación laboral con el sector público o privado en investigadores del área de Ciencias Médicas y de la Salud con filiación universitaria.

2.3. Población y muestra

2.3.1. Población

Investigadores RENACYT con filiación universitaria, y que correspondan al área OCDE de ciencias médicas y de la salud, durante enero del 2021 a octubre del 2023.

2.3.2. Criterio de selección

Criterios de inclusión:

- Registros de investigador RENACYT en el área OCDE de Ciencias Médicas y de la Salud con filiación universitaria.
- Registros de investigador RENACYT durante enero del 2021 a octubre del 2023

Criterios de exclusión:

- Registros de investigador RENACYT que no cuente con la información disponible de acceso público en la plataforma oficial de CONCYTEC.

2.3.3. Muestra y tipo de muestreo

Ingresaron al estudio 1340 registros de investigadores RENACYT con filiación universitaria, y que correspondieron al área OCDE de ciencias médicas y de la salud, entre enero del 2021 a octubre del 2023. No se aplicó muestreo, ingresaron todos aquellos que cumplieron con los criterios de selección.

2.4. Variables

El presente estudio consideró un conjunto estructurado de variables extraídas de fuentes de acceso público (RENACYT y CTI Vitae), clasificadas en: (i) variables independientes (predictoras) y (ii) variables de desenlace (dependientes). La construcción de las variables de producción de investigaciones se realizó respetando el principio de temporalidad, evitando la inclusión de información generada con posterioridad al desenlace de interés. Este criterio permitió construir variables predictoras respetando la direccionalidad temporal en la producción de investigaciones hasta el desenlace de interés.

2.4.1. Variables independientes (predictoras)

Las variables predictoras fueron organizadas en dos grupos según su disponibilidad temporal: variables no transformables (sin registro de fecha) y variables transformables (acumulativas hasta la fecha del desenlace).

a) Variables ajustadas temporalmente

Estas variables fueron calculadas hasta la fecha en que el individuo cumplió la condición de desenlace de interés de vinculación laboral, garantizando un marco temporal: Número de publicaciones científicas (según cuartil SCImago en Q1, Q2, Q3, Q4 y sin cuartil (SC), Número de tesis asesoradas (por nivel de formación: Licenciatura, Segunda Especialidad, Maestría, Doctorado), Número de proyectos de investigación con financiamiento externo reportados.

b) Variables sin ajuste temporal

Estas variables participaron en un subanálisis y se incluyeron como constantes para todos los modelos, al no contar con fechas específicas que permitan establecer un corte temporal hasta el evento: Grado académico más alto alcanzado (variable ordinal codificada como: Bachiller, Licenciado o Título profesional, Segunda especialidad, Maestría, Doctorado), nivel de investigador RENACYT (clasificado como distinguido y niveles de 1 al 7), Índice H (H-index), idioma inglés, número de idiomas distintos del español o inglés, número de desarrollos industriales registrados, número de patentes reportadas.

2.4.2. Variable dependiente (desenlace)

El estudio consideró un desenlace binario independiente, que refleja la vinculación laboral de los investigadores registrados en RENACYT. La categorización se basó en el etiquetado manual del tipo de institución consignada en el CTI Vitae, clasificando a cada investigador según el patrón de filiación institucional vigente. El desenlace fue codificado como una variable dicotómica (1 = cumple la condición de Vinculación con academia y sector público o privado; 0 = solo cumple condición de vinculación con la academia):

- **Vinculación solo con la academia:** Afiliación institucional solamente a universidades, sin vínculos con entidades del sector público ni con empresas privadas.
- **Vinculación con academia y sector público o privado:** Afiliación simultánea con una universidad y con entidades del sector público o empresas privadas (empresas, centros de I+D corporativos, clínicas privadas, etc.).

2.5. Procedimientos

El presente estudio se llevó a cabo siguiendo una metodología rigurosa de recolección, procesamiento y análisis de datos, conforme a los estándares éticos y científicos aplicables. Tras la aprobación del protocolo por parte del Comité Institucional de Ética en Investigación de la Universidad Peruana Cayetano Heredia, se procedió a la recolección de datos de acceso público a partir de la plataforma del RENACYT, administrados por el CONCYTEC.

Se accedió a la plataforma oficial de RENACYT (<https://servicio-renacyt.concytec.gob.pe/>) y a la herramienta de búsqueda de investigadores (<https://servicio-renacyt.concytec.gob.pe/busqueda-de-investigadores/>), seleccionando el filtro correspondiente a la aplicación del reglamento RENACYT-2021. A partir de esta fuente, se identificaron e incluyeron los perfiles de los investigadores clasificados en el área OCDE de Ciencias Médicas y de la Salud, cuyos registros se encontraban disponibles entre enero de 2021 y octubre de 2023.

Para cada investigador se accedió a su ficha individual de RENACYT y al correspondiente registro curricular en la plataforma CTI Vitae. La extracción de datos fue realizada de forma manual y sistemática, construyéndose una base de datos en formato Excel (xlsx) que incluyó las variables definidas en el protocolo, tanto de tipo cualitativo como cuantitativo. A partir de dicha base, se procedió al etiquetado manual del tipo de vinculación laboral (según afiliación con universidad, empresa y/o sector público), así como la clasificación del tipo de universidad (pública o privada) y agrupación de cargos según jerarquía funcional y tipo de institución.

El procesamiento, análisis y modelado de los datos se realizó en la plataforma colaborativa Google Colab, empleando el lenguaje de programación Python (versión 3.7) en entornos de desarrollo tipo Jupyter Notebooks. La base de datos se dividió aleatoriamente en dos subconjuntos: un 80% de los casos para entrenamiento (training set) y un 20% para validación (test set). Esta división se realizó asegurando la aleatoriedad estratificada por desenlace, para mantener representatividad en cada subconjunto. En la figura 1 se observa el flujo de procesamiento y como se procedió para la aplicación de algoritmos de ML en el modelado de los datos.

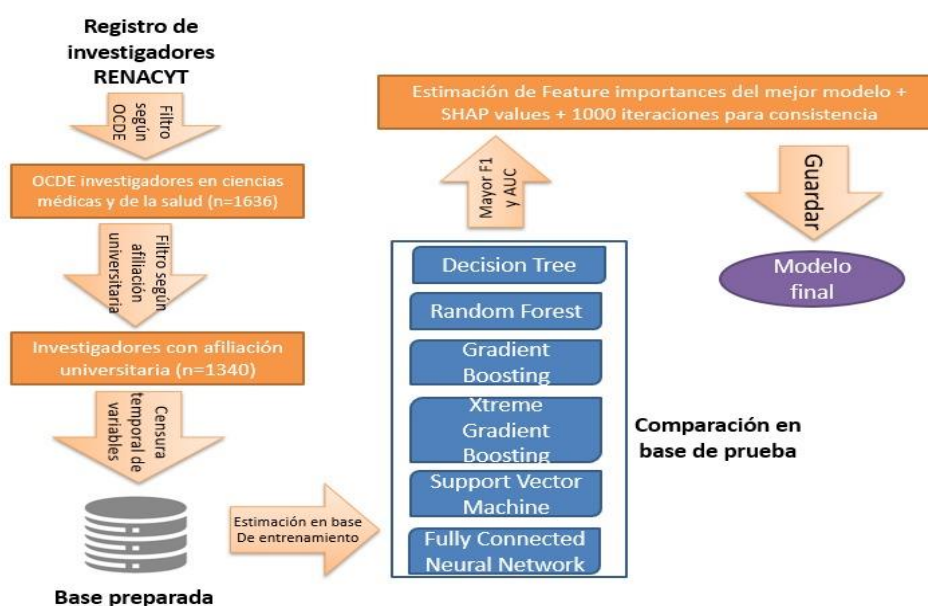
Se diseñó como variable desenlace (outcome): i) vinculación exclusiva con la academia, ii) vinculación con academia y sector público o sector privado. Para estimar este desenlace se construyeron variables predictoras considerando estrictamente el principio de temporalidad, es decir, se excluyeron del modelo información de variables cuya ocurrencia fuera posterior a la fecha en que el sujeto cumplió con el desenlace, con el fin de evitar fuga de información (*data leakage*) y sobreestimación del rendimiento del modelo. En casos donde no se alcanzó el desenlace ($Y=0$), se consideraron todos los datos disponibles hasta la fecha de corte.

Para un análisis adicional y exploratorio del desenlace de interés, se incluyeron las variables predictoras no transformables (por falta de información temporal disponible), como el grado académico más alto, nivel de RENACYT, índice H, idiomas, número de desarrollos industriales y patentes, así como variables transformables, entre ellas número de publicaciones por cuartil (Q1–Q4, sin

cuartil), tesis asesoradas por nivel de formación (licenciatura, segunda especialidad, maestría y doctorado), y proyectos financiados. La adecuación y limpieza de datos se realizó siguiendo criterios de integridad, consistencia y codificación uniforme.

Finalmente, se entrenaron y validaron múltiples algoritmos de ML, incluyendo regresión logística, árboles de decisión, random forest, support vector machines (SVM), gradient boosting, XGBoost y arquitecturas de redes neuronales completamente conectadas (fully connected neural networks). La validación del rendimiento se realizó utilizando métricas como exactitud (accuracy), sensibilidad, especificidad, F1-score, valor predictivo positivo (VPP), valor predictivo negativo (VPN) y área bajo la curva (AUC). Para el modelo de mejor desempeño, se evaluó la importancia relativa de las variables predictoras mediante técnicas de *feature importance* específicas del algoritmo (por ejemplo, ganancia acumulada en árboles de decisión), y se emplearon valores SHAP (*SHapley Additive exPlanations*) para interpretar el impacto individual de cada variable sobre el resultado del modelo, fortaleciendo la comprensión e interpretabilidad del sistema predictivo desarrollado.

Figura 1. Flujo de procesamiento de datos y técnicas utilizadas en el modelado mediante machine learning.



RENACYT: Registro Nacional Científico, Tecnológico y de Innovación Tecnológica; OCDE: Organización para la Cooperación y el Desarrollo Económicos; AUC: Área bajo la curva ROC;

F1: F1-score (media armónica entre precisión y sensibilidad); SHAP: SHapley Additive exPlanations (método para explicar predicciones de modelos de machine learning). La censura temporal de variables se refiere al uso exclusivo de información disponible hasta el momento del evento de interés (número de publicaciones científicas, tipo de cuartil en SCOPUS, número de tesis dirigidas según grado académico, número de proyectos con financiamiento). XGBoost: Xtreme Gradient Boosting. Fully Connected Neural Network: red neuronal multicapa completamente conectada.

2.6. Plan de Análisis

El análisis estadístico se estructuró en dos fases principales: (1) análisis descriptivo de las características de la población de estudio, y (2) entrenamiento, validación e interpretación de modelos predictivos mediante algoritmos de ML.

2.6.1. Análisis Descriptivo

Inicialmente, se realizó una caracterización general de los investigadores RENACYT del área de Ciencias Médicas y de la Salud registrados entre enero de 2021 y octubre de 2023. Para las variables categóricas se calcularon frecuencias absolutas y proporciones relativas. Para las variables numéricas se reportaron medidas de tendencia central y dispersión: mediana y rangos intercuartílicos (RIQ).

2.6.2. Construcción de Modelos Predictivos

La base de datos fue aleatoriamente dividida en dos subconjuntos: un 80% de la muestra para entrenamiento (train set) y el 20% restante para validación (test set). La aleatorización fue estratificada según la variable de desenlace, a fin de preservar la proporción de clases en ambos subconjuntos.

En la base de entrenamiento se desarrollaron modelos de ML para estimar el desenlace de interés, los algoritmos obtenidos fueron medidos para evidenciar sus métricas en la muestra de validación. Los modelos se entrenaron utilizando los siguientes algoritmos de ML: Regresión logística, Árbol de decisión (Decision Tree Classifier), Bosques aleatorios (Random Forest), Gradient Boosting y Xtreme Gradient Boosting (XGBoost), Máquinas de vectores de soporte (Support Vector Machine, SVM), Redes neuronales completamente conectadas (Fully Connected Neural Networks).

Se priorizó el uso de métricas robustas para evaluar el rendimiento de los modelos en la muestra de validación, incluyendo: Accuracy (Exactitud global), Precisión

(Valor predictivo positivo), Recall (Sensibilidad), F1-score, Especificidad, Valor predictivo negativo (VPN), Área bajo la curva ROC (AUC-ROC). Además, se realizó un análisis exploratorio adicional que incluía tanto las variables con ajuste temporal como variables sin censura temporal, con la finalidad de comparar el poder predictivo de modelos.

En los modelos de ML con mejor desempeño, con las variables de ajuste temporal, se realizó un análisis de interpretabilidad para identificar las variables predictoras más influyentes. Esto se efectuó mediante: Importancia de características (Feature Importance) para evaluar la contribución relativa de cada variable predictora a la clasificación del modelo, y mediante SHAP values (SHapley Additive exPlanations) con el propósito de estimar el impacto individual de cada variable en la predicción del modelo, proporcionando explicabilidad de los resultados.

Así mismo, para asegurar la consistencia y estabilidad de los valores de área bajo la curva ROC y F1 score obtenidos del modelo de ML con mejor desempeño, se realizaron aleatorizaciones de las muestras de entrenamiento y prueba, generando conjuntos de datos mediante aleatorización repetida en 1,000 iteraciones para su evaluación.

Todas las fases de análisis fueron desarrolladas utilizando Python 3.7 en la plataforma Google Colab, con bibliotecas especializadas en ciencia de datos. Los scripts de análisis fueron documentados para garantizar su trazabilidad, reproducibilidad y futura escalabilidad.

2.7. Consideraciones éticas

El presente estudio se desarrolló respetando los principios éticos fundamentales de la investigación científica, conforme a las disposiciones de la Declaración de Helsinki, las normas del Consejo de Organizaciones Internacionales de las Ciencias Médicas (CIOMS), y la normativa nacional peruana sobre ética en investigación con seres humanos. Si bien el estudio no implicó contacto directo con participantes ni intervención sobre personas, se aplicaron criterios de ética en el manejo de datos personales, análisis y difusión responsable de resultados. Particularmente, se cumplieron aspectos relacionados con la revisión ética de estudios observacionales, protección de la privacidad, uso ético de datos secundarios y aplicación transparente y no discriminatoria de algoritmos de inteligencia artificial.

Previo al inicio de la recolección y procesamiento de datos, el protocolo de investigación fue evaluado y aprobado por el Comité Institucional de Ética en Investigación de la Universidad Peruana Cayetano Heredia, bajo el principio de respeto a la dignidad, privacidad y usando de forma responsable la información que está registrada en plataformas públicas.

Los datos utilizados fueron obtenidos exclusivamente de fuentes de acceso público: el Registro Nacional Científico, Tecnológico y de Innovación Tecnológica (RENACYT) y la plataforma CTI Vitae, ambas pertenecientes al CONCYTEC del Perú. En concordancia con la Ley N.º 29733 – Ley de Protección de Datos Personales, no se accedió, almacenó ni analizó información que permita identificar directamente a las personas, garantizando así el anonimato de los investigadores registrados. La recolección de datos fue realizada manualmente desde la interfaz pública disponible en la web oficial de CONCYTEC, respetando los términos y condiciones de uso establecidos en dichas plataformas, asegurando el cumplimiento con los principios de legalidad y proporcionalidad en la recolección de información.

Adicionalmente, se tomaron medidas para proteger la confidencialidad y seguridad de los datos durante todas las etapas del estudio. La base de datos fue almacenada en entorno controlado y protegidos con credenciales de acceso únicamente por el investigador del estudio. No se compartieron bases de datos crudas fuera del equipo de investigación autorizado y se garantizó la integridad y trazabilidad del tratamiento de la información.

Los resultados obtenidos del presente estudio se presentan exclusivamente en forma agregada y estadística, sin revelar nombres, ni ninguna información que permita la identificación de personas. Asimismo, se declara el compromiso ético de promover su aplicación únicamente para fines de mejora de la política pública y gestión del talento humano científico. Se considera que el presente estudio cumple con los principios de beneficencia, no maleficencia, justicia y respeto por las personas, aplicados al contexto de investigación secundaria con datos públicos y análisis computacional.

III. RESULTADOS

3.1. Descripción de la muestra

La muestra estuvo conformada por 1,340 investigadores registrados en RENACYT que reportaron vinculación académica, cuya clasificación principal correspondía al área OCDE de Ciencias Médicas y de la Salud. Esta categoría agrupó a profesionales que desarrollan investigación en campos como medicina básica, clínica, salud pública, epidemiología, nutrición, biotecnología médica, entre otros. Del total, el 66.9 % (n = 896) reportó una doble vinculación laboral con instituciones académicas y con entidades del sector público o privado, mientras que el 33.1 % (n = 444) indicó afiliación solo al ámbito académico.

Descripción profesional y laboral:

En la tabla 1 se observan los hallazgos encontrados en relación con la descripción de formación profesional y laboral. Respecto a la formación profesional sanitaria, la mayoría correspondía a médicos (27.6 %), seguidos por biólogos (11.5 %), tecnólogos médicos (3.4 %), psicólogos (2.9 %) y obstetras (2.2 %). El 52.4 % restante agrupó a profesionales de otras disciplinas vinculadas al área de la salud. En relación con el grado académico más alto alcanzado, el 46.5 % poseía grado de doctor, el 35.6 % contaba con maestría, el 17.3 % con licenciatura o segunda especialidad, y el 0.6 % con grado de bachiller. En cuanto al perfil jerárquico reportado, el 17.56% desempeñaba cargos de jefatura, coordinación, dirección o gerencia, mientras que el 3.2 % se identificó como analista o especialista. Respecto a la condición contractual, el 42.2 % indicó pertenecer a la modalidad de contratado, el 37.5 % a régimen ordinario, y el 20.3 % a otras formas contractuales. Así mismo, el 92.5 % de los participantes reportó dominio del idioma inglés, frente al 7.5 % que declaró no poseerlo.

Descripción de producción científica y académica:

En la tabla 1 se observan los hallazgos encontrados en relación con la descripción de producción científica y académica. El H-index presentó una mediana de 2 con un rango intercuartílico (RIQ) de 0 a 4. Ello indicó que, aunque el 50 % de los investigadores no superó las dos citas ponderadas, existe un sector con un impacto moderado (percentil 75 $\approx$ 4). En la clasificación RENACYT, la mediana fue el nivel 6 (RIQ: del nivel 4 al 7), ubicando a la mayoría en niveles intermedios

y bajos. En la producción tecnológica (desarrollos industriales y patentes) la mediana y el RIQ reportaron ausencia de registros en toda la distribución.

Para la producción científica según cuartil Scopus se observó una mediana de 0 en Q1–Q4, con RIQ de 0–1. Esto implica que al menos el 50 % de la muestra no publicó en esas categorías, pero hasta 25 % reportó al menos un artículo por cuartil, reflejando actividad limitada en revistas indexadas, en la mayoría de los investigadores. Por otro lado, en las publicaciones en revistas Scopus sin cuartil al menos el 25 % registra hasta cinco artículos, sugiriendo mayor accesibilidad a estos medios en relación con las revistas de mayor cuartil. Así mismo, la mediana de proyectos financiados fue 0, con un rango intercuartílico también de 0 a 0, lo que indica que al menos el 75% de los investigadores no reportó participación en proyectos con financiamiento competitivo durante el periodo evaluado.

La dirección de tesis fue limitada. En licenciatura, la mediana fue 0 (RIQ: 0–1), evidenciando que solo el 25% asesoró al menos una tesis; en maestría, doctorado y segunda especialidad fue incluso más limitado.

Tabla 1. Caracterización profesional, académica y productiva de investigadores RENACYT del área de Ciencias Médicas y de la Salud afiliados a instituciones académicas, con o sin doble vinculación al sector público o privado.

Categoría	Subcategoría	Total=1340	Porcentaje
Grado académico			
	Bachiller	8	0.60%
	Licenciado/Titulado/Segunda especialidad	232	17.31%
	Magíster	477	35.60%
	Doctor	623	46.49%
H index*		2	(0 - 4)
Nivel RENACYT*		6	(4 - 7)
Inglés	No	101	7.54%
	Si	1239	92.46%
Idiomas distintos al español o inglés*		1	(0 - 1)

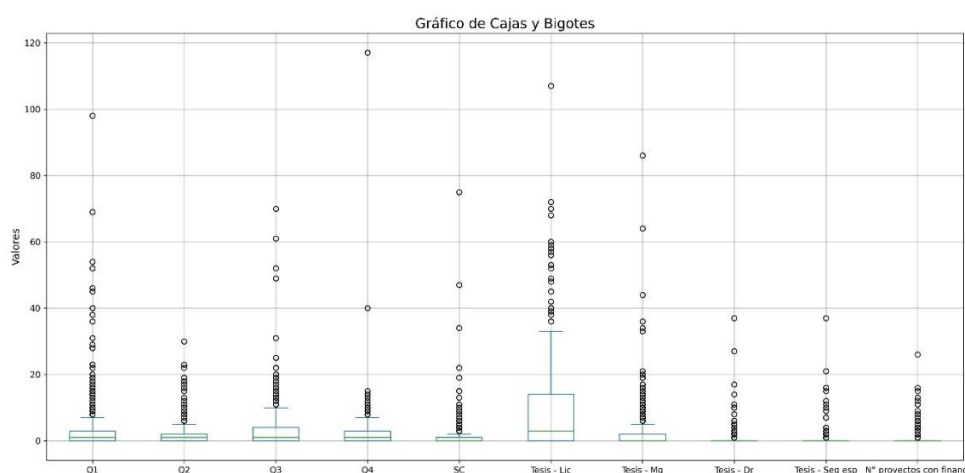
Desarrollos industriales*	0	(0 - 0)
Patentes*	0	(0 - 0)
Publicaciones *		
Primer cuartil	0	(0 - 1)
Segundo cuartil	0	(0 - 1)
Tercer cuartil	0	(0 - 1)
Cuarto cuartil	0	(0 - 1)
Sin cuartil	0	(0 - 5)
Perfil laboral		
Solo docente	444	33.10%
Analista / especialista	43	3.21%
Coordinación	49	3.66%
Jefatura	80	5.97%
Dirección / Gerencia	105	7.84%
Otros perfiles	619	46.23%
Asesoría de tesis*		
Licenciatura	0	(0 - 1)
Maestría	0	(0 - 0)
Doctorado	0	(0 - 0)
Segunda especialidad	0	(0 - 0)
Tipo de contrato		
Contratado	565	42.16%
Ordinario	503	37.54%
Otros	272	20.30%
Proyectos financiados*	0	(0 - 0)
Profesión		
Medico	370	27.61%
Biólogo	154	11.49%
Tecnólogo médico	46	3.43%
Psicólogo	39	2.91%
Obstetra	29	2.16%
Otro	702	52.39%

*Variables numéricas expresadas en mediana y rango intercuartílico (RIQ)

3.2. Descripción de la producción científica y académica de investigadores RENACYT según vinculación laboral

La figura 1 muestra la producción académica y científica de investigadores RENACYT con afiliación solamente académica. Se observó una productividad general baja, con medianas de cero en la mayoría de las variables y alta frecuencia de valores atípicos. Las publicaciones Scopus sin cuartil (SC) y en revistas Q1 presentaron mayor dispersión (mediana: 2 y 1, respectivamente), mientras que la asesoría de tesis se concentró en el nivel de licenciatura (mediana: 5; RIQ: 1–14). La participación en tesis de posgrado y en proyectos con financiamiento fue escasa, con medianas nulas.

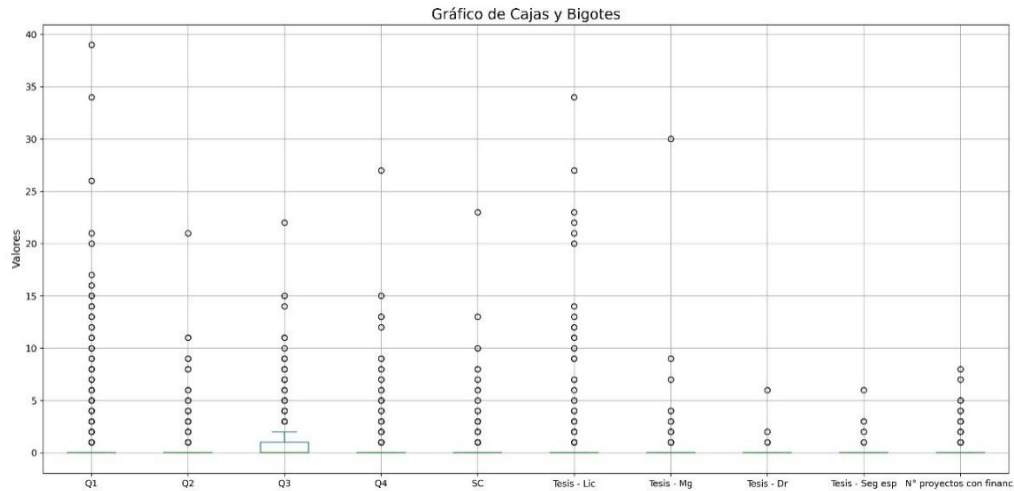
Figura 2. Distribución de la producción académica y científica de investigadores RENACYT con vinculación solamente académica.



Las variables "Tesis – Lic.", "Mg", "Dr." y "Seg esp" corresponden al número de tesis asesoradas en los niveles de licenciatura, maestría, doctorado y segunda especialidad, respectivamente. Q1–Q4 y SC indican el número de publicaciones en revistas científicas clasificadas por cuartil. "N° proyectos con financ" corresponde al número de proyectos con financiamiento.

La figura 2 corresponde a investigadores con vinculación laboral académica y con el sector público o privado, considerando únicamente la producción previa a la obtención de ambas filiaciones. La actividad científica en este periodo fue particularmente baja, con distribuciones concentradas en cero en casi todas las variables. Solo las publicaciones Q3 presentaron una mediana distinta de cero (mediana: 1; RIQ: 0–2). Las asesorías de tesis y los proyectos con financiamiento también fueron mínimos, indicando una limitada productividad previa a la expansión institucional.

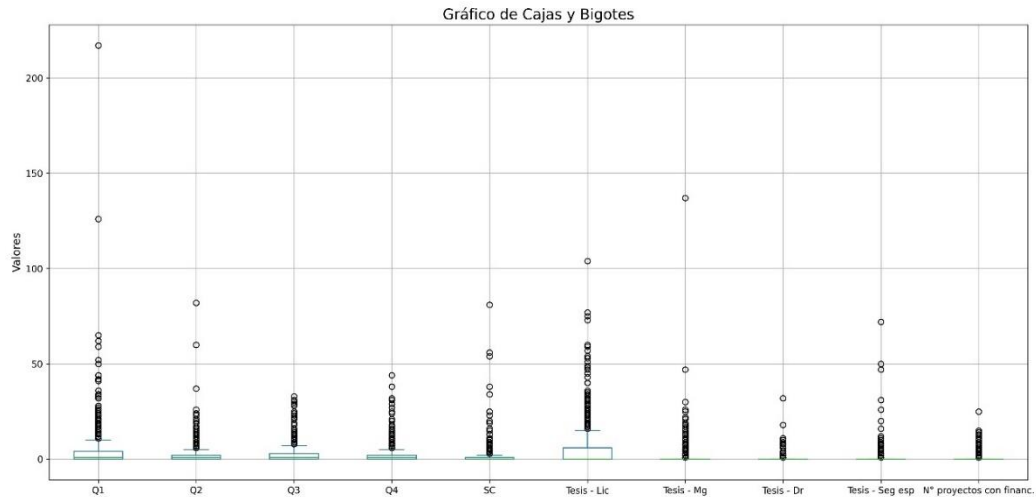
Figura 3. Distribución de la producción académica y científica de investigadores RENACYT antes de tener vinculación académica y con el sector público o privado.



Las variables "Tesis – Lic.", "Mg", "Dr." y "Seg esp" corresponden al número de tesis asesoradas en los niveles de licenciatura, maestría, doctorado y segunda especialidad, respectivamente. Q1–Q4 y SC indican el número de publicaciones en revistas científicas clasificadas por cuartil. “N° proyectos con financ” corresponde al número de proyectos con financiamiento.

En contraste, la figura 3 corresponde a investigadores con vinculación laboral académica y con el sector público o privado, y presenta la producción posterior a la doble afiliación. Se observó un incremento significativo en todas las métricas. Las publicaciones en revistas Q1 mostraron una mediana de 2 (RIQ: 0–5) con valores extremos, y se registró mayor participación en asesoría de tesis, especialmente en licenciatura (mediana: 8; RIQ: 3–20). Aunque las medianas en niveles de posgrado y proyectos financiados permanecieron con una mediana de cero, la frecuencia y magnitud de los valores atípicos aumentaron, sugiriendo un mayor dinamismo investigador tras la incorporación simultánea a ambos sectores.

Figura 4. Distribución de la producción académica y científica de investigadores RENACYT después de la vinculación académica y con el sector público o privado.



Las variables "Tesis – Lic.", "Mg", "Dr." y "Seg esp" corresponden al número de tesis asesoradas en los niveles de licenciatura, maestría, doctorado y segunda especialidad, respectivamente. Q1–Q4 y SC indican el número de publicaciones en revistas científicas clasificadas por cuartil. "N° proyectos con financ" corresponde al número de proyectos con financiamiento.

En conjunto, los resultados evidencian una marcada heterogeneidad en la trayectoria investigadora. La vinculación laboral académica y con el sector público o privado correspondió con un mayor volumen y diversidad de producción académica y científica, especialmente tras su afiliación; sin embargo, esta mayor productividad no es generalizada, sino que se concentra en un grupo menor de investigadores con mayor productividad.

3.3. Modelos predictivos con censura temporal (modelo principal)

Este análisis incluyó únicamente variables cuya trazabilidad temporal permitió censurarlas hasta la fecha de ocurrencia del desenlace, garantizando que ninguna información posterior al evento de interés haya sido incluida. El desenlace fue definido como una variable dicotómica: 1 = vinculación simultánea con academia y sector público o privado, y 0 = vinculación exclusiva con la academia.

Se entrenaron múltiples algoritmos supervisados sobre un conjunto de entrenamiento (80%) y se validaron sobre un conjunto de prueba (20%), empleando únicamente variables censuradas de forma adecuada.

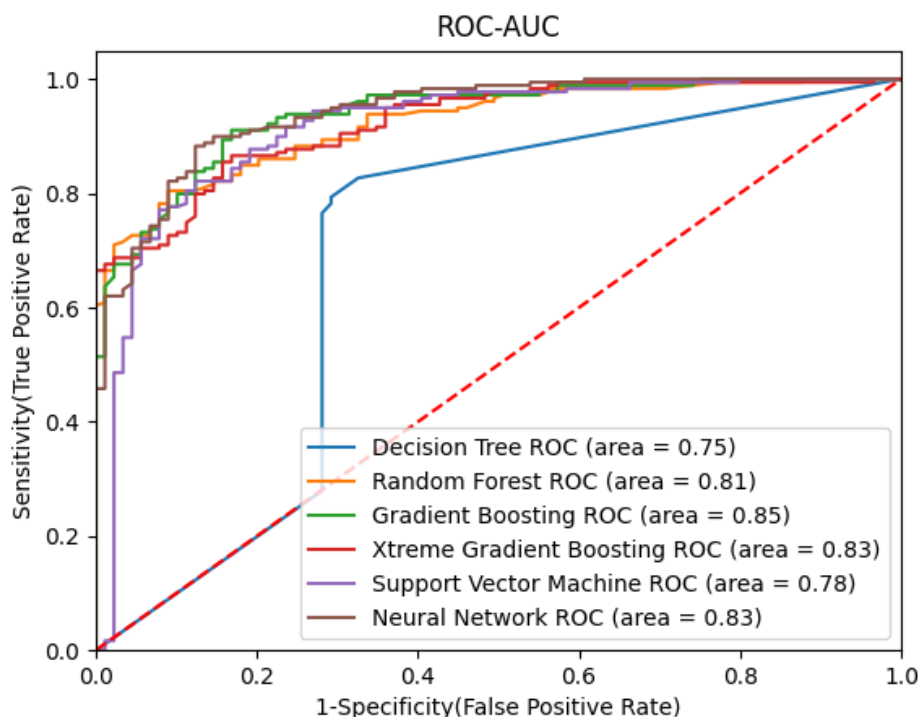
3.3.1. Análisis de Curvas ROC

La evaluación del rendimiento discriminativo de los modelos entrenados mediante curvas ROC (Figura 4) evidenció que el algoritmo Gradient Boosting alcanzó el mejor desempeño global, con un área bajo la curva (AUC) de 0.85, superando a los demás modelos evaluados.

El modelo Xtreme Gradient Boosting (XGBoost) y la Red Neuronal Totalmente Conectada (Neural Network) mostraron desempeños similares, ambos con un AUC de 0.83, seguidos por Random Forest con un AUC de 0.81. Los modelos Support Vector Machine (SVM) y Árbol de Decisión obtuvieron desempeños más bajos, con AUC de 0.78 y 0.75, respectivamente.

La forma y altura de la curva ROC del modelo Gradient Boosting indican un excelente equilibrio entre sensibilidad y especificidad, manteniendo una alta tasa de verdaderos positivos a lo largo de todo el rango de falsos positivos. Este comportamiento sugiere que dicho modelo tiene una alta capacidad para distinguir entre investigadores con vinculación exclusivamente académica y aquellos con doble afiliación (academia y sector público o privado).

Figura 5. Curvas ROC comparativas de modelos de machine learning para la predicción de vinculación laboral utilizando variables censuradas temporalmente.



Curvas ROC (Receiver Operating Characteristic) de los modelos de machine learning entrenados para predecir la probabilidad de vinculación laboral simultánea con la academia y el sector público o privado. El eje X representa la tasa de falsos positivos (1 - especificidad), y el eje Y representa la tasa de verdaderos positivos (sensibilidad). La línea diagonal discontinua corresponde a un modelo sin capacidad discriminativa (AUC = 0.50).

3.3.2. Métricas de rendimiento del modelo Gradient Boosting y comparación con otros algoritmos.

Se evaluó el desempeño predictivo de múltiples algoritmos de ML para predecir la doble afiliación laboral (academia y sector público o privado), utilizando como entradas variables censuradas temporalmente.

Entre los modelos entrenados, el Gradient Boosting Classifier alcanzó el mejor rendimiento global, con un AUC-ROC de 0.85, una exactitud (accuracy) de 0.87, precisión de 0.90, sensibilidad de 0.91, especificidad de 0.79, F1-score de 0.87 y un valor predictivo positivo (VPP) de 0.81. Estos valores reflejan un balance óptimo entre la capacidad discriminativa del modelo y su aplicabilidad para identificar correctamente casos positivos (doble afiliación) sin comprometer la precisión ni la especificidad.

Le siguieron en rendimiento el MLP Classifier (red neuronal completamente conectada) con un AUC de 0.83, exactitud de 0.87 y F1-score de 0.86, y el XGBClassifier, también con AUC de 0.83 y F1-score de 0.84. Otros algoritmos como Random Forest y SVC (Support Vector Classifier) mostraron desempeños aceptables, aunque con menor robustez: AUCs de 0.81 y 0.78, respectivamente. El Árbol de Decisión fue el modelo con menor rendimiento general (AUC = 0.75, F1 = 0.77).

Cabe destacar que el modelo Gradient Boosting no solo obtuvo la mayor área bajo la curva ROC, sino también el mejor equilibrio entre sensibilidad (0.91) y especificidad (0.79), lo cual es especialmente relevante en contextos donde la correcta identificación de casos con doble vinculación institucional resulta crítica para informar decisiones de política científica y asignación de recursos.

Tabla 2. Comparación de métricas de rendimiento de modelos para la predicción de vinculación laboral utilizando variables censuradas temporalmente.

Modelo	Exactitud	PPV	Especificidad	Sensibilidad	F1-score	AUC	FPV
Decision Tree Classifier	0.76	0.85	0.71	0.79	0.77	0.75	0.63
Random Forest Classifier	0.83	0.88	0.75	0.87	0.83	0.81	0.74
Gradient Boosting Classifier	0.87	0.9	0.79	0.91	0.87	0.85	0.81
XGB Classifier	0.84	0.89	0.79	0.87	0.84	0.83	0.74
SVC	0.84	0.83	0.6	0.96	0.83	0.78	0.88
MLP Classifier	0.87	0.87	0.71	0.94	0.86	0.83	0.86

PPV: valor predictivo positivo; AUC: área bajo la curva ROC; FPV: valor predictivo negativo; F1-score: medida armónica entre precisión y sensibilidad; SVC: Support Vector Classifier; MLP: red neuronal multicapa.

Las métricas fueron calculadas sobre el conjunto de validación, utilizando únicamente variables predictoras con registro de fecha, las cuales fueron censuradas temporalmente hasta el momento de ocurrencia del desenlace.

3.3.3. Feature importance del modelo con mayor rendimiento

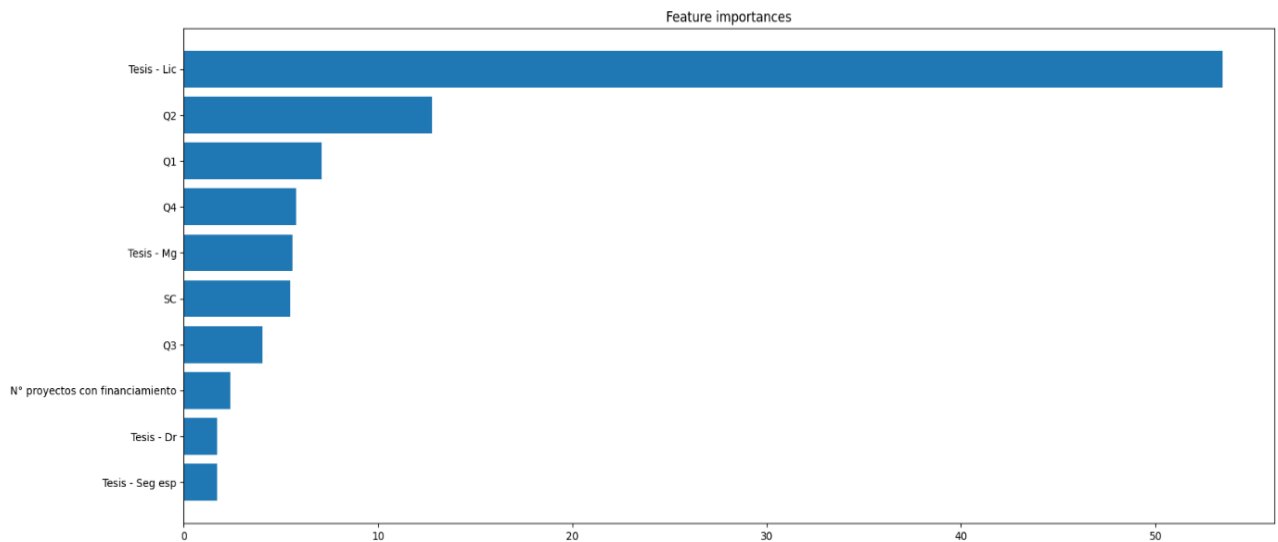
La interpretación de la importancia de características en el modelo Gradient Boosting se realizó utilizando la métrica de ganancia acumulada (total gain), la cual cuantifica la contribución relativa de cada variable a la mejora del rendimiento del modelo. Específicamente, esta métrica refleja la suma de las reducciones en la función de pérdida (log-loss) obtenidas en todos los nodos de decisión en los que una variable fue utilizada para dividir el árbol. Feature importance estima el efecto de cada variable de forma marginal, es decir, sin considerar explícitamente las interacciones con otras variables.

La Figura 5 muestra la jerarquía de importancia de las variables incluidas en el modelo principal, el cual fue entrenado exclusivamente con variables censuradas hasta la fecha del desenlace. La variable más influyente fue el número de tesis de licenciatura asesoradas, que representó más del 50 % de la ganancia total del modelo, indicando una fuerte capacidad de discriminación para diferenciar entre investigadores con y sin vinculación combinada.

Le siguieron en importancia las publicaciones en revistas Q2, Q1 y Q4, junto con las tesis de maestría asesoradas, las publicaciones sin cuartil (SC), y el número de tesis en otros niveles (doctorado, segunda especialidad). El número de proyectos de investigación con financiamiento tuvo una contribución menor, aunque no despreciable.

Este patrón sugiere que las variables relacionadas con productividad científica y actividad formativa, particularmente en el pregrado, fueron las que más ayudaron al modelo a separar vinculación combinada de la vinculación exclusivamente académica. Es importante señalar que esta métrica de importancia de variables no permite inferir la dirección del efecto de cada variable sobre el desenlace, sino que permite evaluar el peso de cada variable con respecto a la capacidad discriminatoria de clasificación del modelo predictivo.

Figura 6. Importancia relativa de las variables predictoras en el modelo Gradient Boosting (modelo principal, con censura temporal).



Las variables "Tesis – Lic", "Mg", "Dr" y "Seg esp" corresponden al número de tesis asesoradas en los niveles de licenciatura, maestría, doctorado y segunda especialidad, respectivamente. Q1–Q4 y SC indican el número de publicaciones en revistas científicas clasificadas por cuartil. “N° proyectos con financiamiento” corresponde al total acumulado hasta la fecha del desenlace. Esta métrica no proporciona información sobre la dirección del efecto de los predictores.

3.3.4. SHAP values del modelo con mayor rendimiento

Para interpretar la dirección y magnitud del impacto de cada variable sobre la predicción del modelo, se aplicó el método de SHAP values (SHapley Additive exPlanations). Los valores SHAP permitieron estimar, para cada observación individual, el efecto marginal de cada variable sobre la salida del modelo, expresado en log-odds o probabilidades ajustadas.

La Figura 6 presenta el summary plot de los valores SHAP para el modelo Gradient Boosting, construido con variables censuradas temporalmente que explican la probabilidad predicha de vinculación simultánea con academia y sector público o privado.

En relación con las variables correspondientes a producción científica por cuartiles, se observaron patrones diferenciados en la contribución de cada tipo de publicación:

- Publicaciones Q1: Aunque algunos valores extremos con alto número de publicaciones Q1 mostraron contribuciones positivas significativas a la predicción del desenlace ($\text{SHAP} > 0$), la mayoría de los puntos rojos se agruparon en la zona de $\text{SHAP} \leq 0$. Esto indica que, en general, un mayor número de publicaciones Q1 no se asocia consistentemente con mayor probabilidad de doble afiliación institucional, y en muchos casos incluso puede contribuir de forma negativa. Este hallazgo sugiere la presencia de interacciones o modulaciones del efecto de Q1 con otras variables del modelo (por ejemplo, grado académico o experiencia en tesis).
- Publicaciones Q2: Se evidenció una distribución centrada en torno a $\text{SHAP} \approx 0$, con predominancia de puntos rojos (valores altos) en la zona negativa. Ello implica que un mayor número de publicaciones en Q2, de forma aislada, no impulsa la predicción hacia la doble afiliación y puede incluso reducirla marginalmente.
- Publicaciones Q3: La distribución fue más equilibrada, con puntos rojos tanto en la zona negativa como positiva de los valores SHAP. Sin embargo, la densidad fue ligeramente mayor en $\text{SHAP} < 0$, lo que sugiere un efecto en gran medida neutro, con escasa direccionalidad independiente hacia el desenlace.
- Publicaciones Q4: Se observó un patrón negativo claro. La mayoría de los puntos rojos se agruparon en $\text{SHAP} \leq 0$, evidenciando que altos valores en esta variable no solo no favorecen, sino que tienden a disminuir la probabilidad de predicción positiva. La capacidad discriminativa de esta variable fue, por tanto, limitada.
- Publicaciones sin cuartil (SC): La tendencia fue marcadamente negativa. La acumulación de puntos rojos en valores $\text{SHAP} \leq 0$ indica que esta variable contribuye consistentemente en contra de la predicción de doble afiliación, sugiriendo que una mayor proporción de publicaciones en revistas no indexadas no es un factor relevante para vinculación multisectorial, al menos desde la perspectiva del modelo.

En el caso del número de tesis asesoradas (de licenciatura, maestría, doctorado y segunda especialidad), se observó que mayores valores se concentraron consistentemente en la zona $SHAP \leq 0$, indicando que, en este modelo, una mayor experiencia en asesoría de tesis se asoció con una menor probabilidad de doble afiliación.

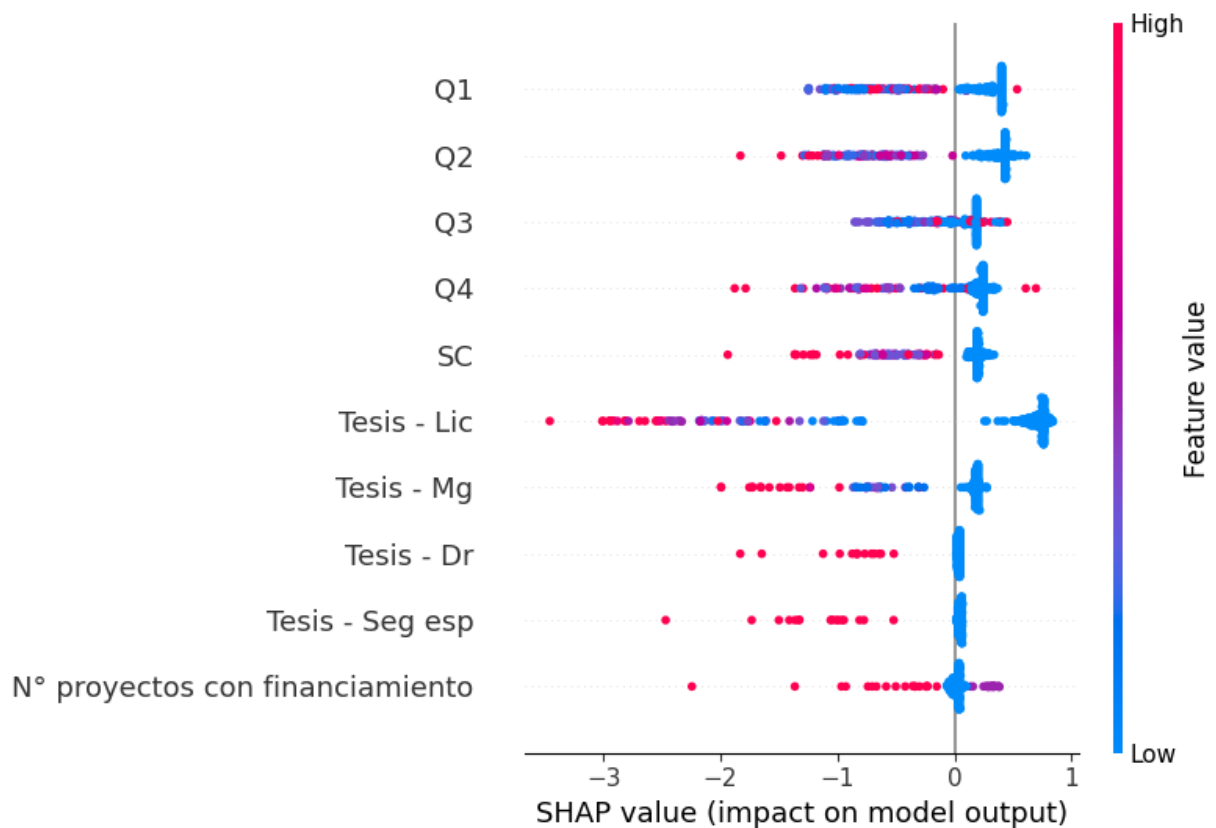
Por otro lado, la variable número de proyectos con financiamiento presentó un patrón más complejo. Los valores bajos se concentraron cerca de $SHAP \approx 0$, lo que sugiere que tener pocos proyectos no influye significativamente en la clasificación del modelo. Sin embargo, se identificaron valores altos (rojos y morados) tanto en la zona $SHAP \leq 0$ como en $SHAP > 0$, reflejando un efecto bidireccional. En algunos casos, un alto número de proyectos se asoció a menor probabilidad de doble vinculación, mientras que, en otros casos, particularmente en combinación con otras variables, contribuyó fuertemente a la predicción positiva.

En conjunto, estos hallazgos sugieren que, si bien la producción científica y la participación en proyectos son predictivamente relevantes, su impacto direccional no es homogéneo y depende del perfil contextual de cada investigador. Esto resalta la importancia de modelos interpretables y del uso de métodos como SHAP para descomponer los efectos individuales de variables en entornos complejos y multivariantes como el que caracteriza al sistema de ciencia y tecnología del Perú.

El orden vertical de las variables en la Figura 6 fue determinado por la importancia global de cada predictor en el modelo, calculado como la media del valor absoluto de los SHAP values en toda la muestra. Este enfoque no evaluó si una variable tiene un efecto positivo o negativo, sino cuánto modifica la salida del modelo en promedio, sin importar la dirección. En consecuencia, las variables ubicadas en la parte superior (como Q1, Q2 y Q3) fueron las que más contribuyeron a la variabilidad de las predicciones, actuando como factores de mayor peso explicativo para la clasificación del tipo de vinculación institucional. Por el contrario, variables como tesis de segunda especialidad o

el número de proyectos con financiamiento, aunque relevantes en ciertos individuos, tuvieron una contribución global más reducida, al presentar valores SHAP generalmente más pequeños o distribuidos con baja dispersión. Este orden complementa la interpretación direccional y condicional ofrecida por los SHAP values, proporcionando una visión jerárquica de las variables más influyentes desde la perspectiva estructural del modelo de ML.

Figura 7. Gráfico resumen de valores SHAP del modelo Gradient Boosting entrenado con variables censuradas temporalmente.



Las variables Q1–Q4 y SC indican el número de publicaciones en revistas científicas clasificadas por cuartil. Tesis – Lic, Mg, Dr y Seg esp corresponden al número de tesis asesoradas en los niveles de licenciatura, maestría, doctorado y segunda especialidad, respectivamente. N° proyectos con financiamiento: corresponde al total acumulado hasta la fecha del desenlace. El eje Y lista las variables predictoras ordenadas por importancia global (impacto promedio absoluto del SHAP value). El eje X muestra el SHAP value individual: cuánto contribuye una observación específica (un investigador) a mover la predicción hacia el desenlace positivo (SHAP > 0) o negativo (SHAP < 0). El color indica el valor real de la variable en cada observación: rojo = alto valor, azul = bajo valor.

3.4. Consistencia del modelo de ML con mejor desempeño

La estabilidad del modelo Gradient Boosting fue evaluada mediante 1,000 repeticiones de entrenamiento y validación aleatorizadas, utilizando como métricas de desempeño el área bajo la curva ROC (AUC-ROC) y el F1-score. En la Figura 7, se observó que los valores de AUC-ROC en la muestra de entrenamiento se mantuvieron elevados y consistentes, con un rango estrecho entre 0.93 y 0.95, mientras que en la muestra de validación presentaron una media inferior (aproximadamente entre 0.80 y 0.85) y mayor variabilidad. En la Figura 8, los valores de F1-score para la muestra de entrenamiento se concentraron en torno a 0.96 con mínima dispersión, mientras que en la validación oscilaron entre 0.86 y 0.92. Estos resultados evidenciaron una alta consistencia del modelo frente a distintas particiones de los datos, sin indicios significativos de sobreajuste, y respaldan su robustez y generalización para tareas de clasificación con estructuras de datos similares.

Figura 8. Distribución de valores del área bajo la curva ROC en repeticiones de entrenamiento y validación del modelo Gradient Boosting.

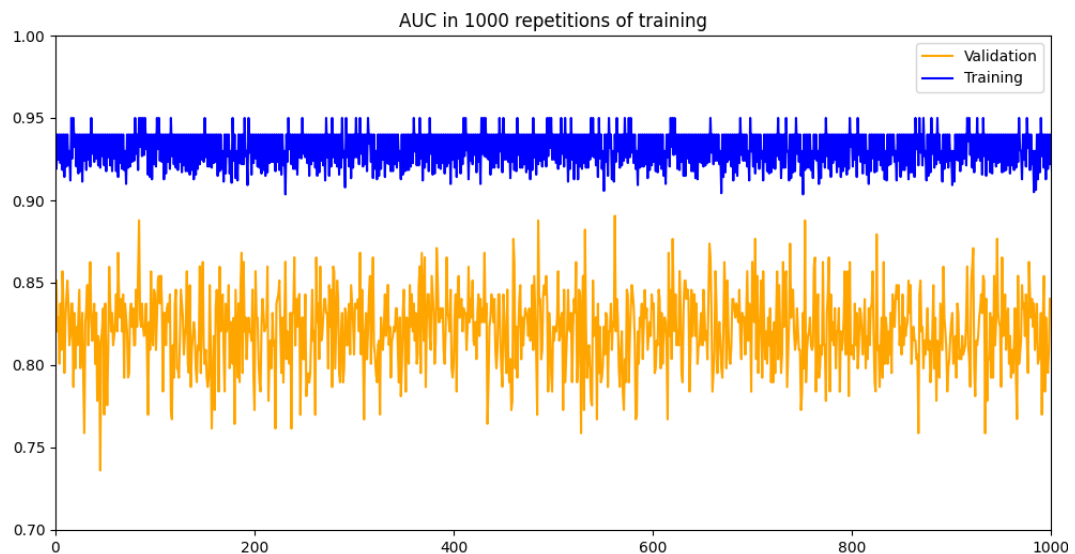
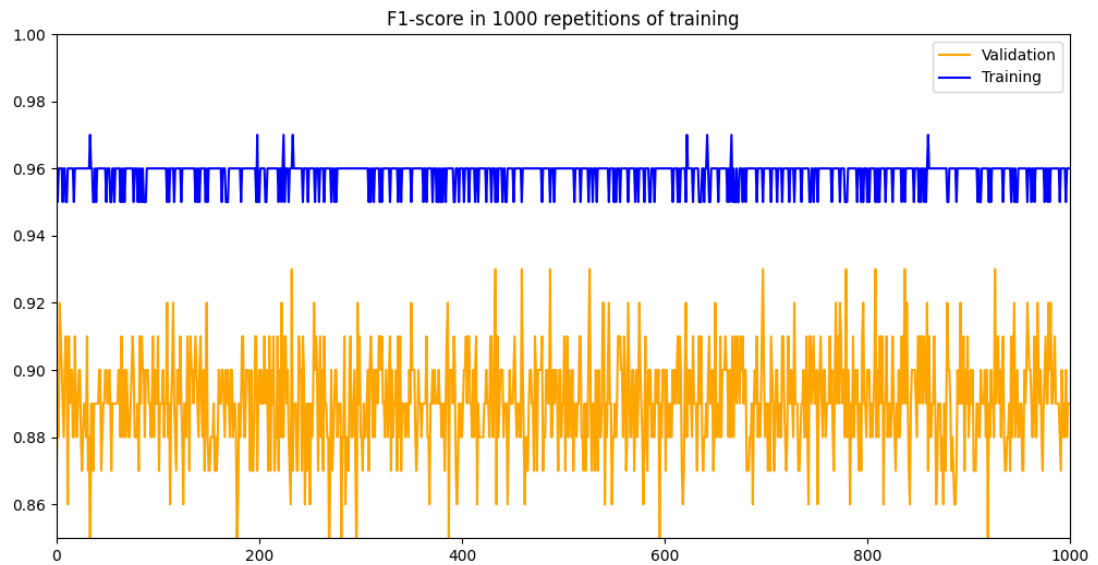


Figura 9. Distribución de valores de F1 score en repeticiones de entrenamiento y validación del modelo Gradient Boosting.



3.5. Análisis exploratorio con variables extendidas no censuradas

Adicionalmente, se entrenó un modelo exploratorio utilizando un conjunto extendido de variables predictoras, que incluyó tanto las variables censuradas por temporalidad como aquellas sin trazabilidad temporal (modelo híbrido). Este enfoque no constituye un análisis de sensibilidad en sentido estricto, sino un modelo complementario orientado a estimar la ganancia marginal de información cuando se permite el uso de variables estáticas o no fechadas al cierre del estudio. Sin embargo, su uso operativo debe balancearse frente al riesgo de incorporar información posterior al evento, lo que limita su aplicabilidad prospectiva.

3.5.1. Análisis de Curvas ROC de modelos

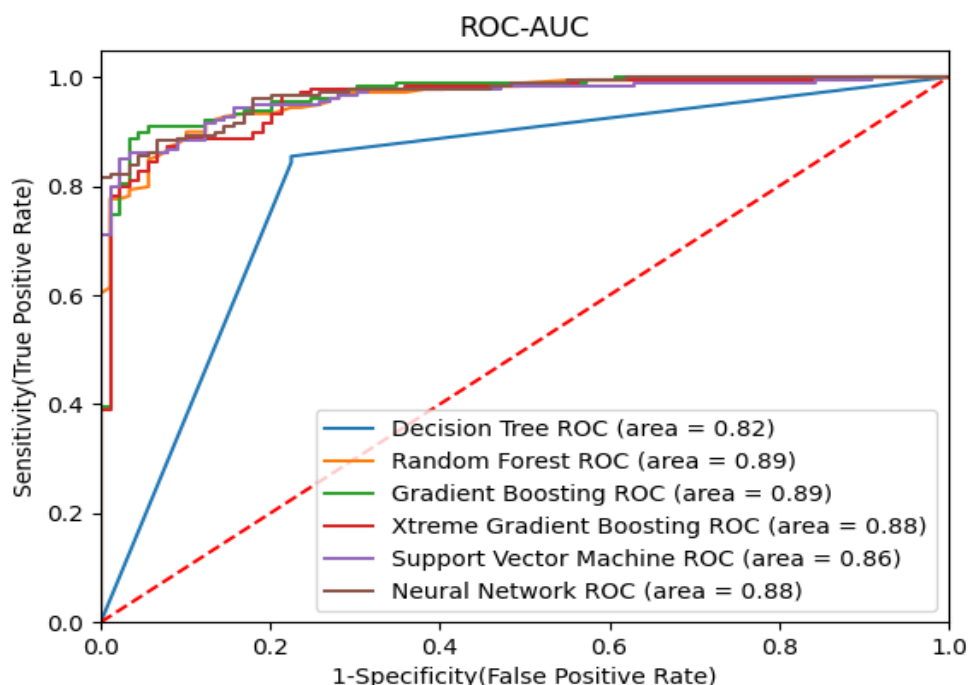
Como parte del análisis exploratorio, se evaluó el desempeño discriminativo de distintos algoritmos de ML entrenados con un conjunto extendido de variables predictoras, que incluyó tanto aquellas con temporalidad censurada como aquellas sin trazabilidad temporal al evento. Este modelo híbrido no constituyó un análisis de sensibilidad en sentido estricto, sino un ejercicio complementario que estimó el comportamiento del modelo en escenarios donde se incluyeron características adicionales disponibles al cierre del estudio.

La Figura 9 muestra las curvas ROC comparadas de seis algoritmos. El rendimiento global, medido mediante el área bajo la curva (AUC-ROC), fue

elevado en todos los casos, superando el umbral de 0.80. El mejor desempeño se observó en los modelos Random Forest y Gradient Boosting, ambos con un AUC de 0.89, seguidos por XGBoost y Red Neuronal Multicapa con un AUC de 0.88, y SVM con un AUC de 0.86. El modelo con menor rendimiento fue el Árbol de Decisión, aunque aún alcanzó un AUC aceptable de 0.82.

Comparado con el modelo principal basado únicamente en variables censuradas (AUC máximo = 0.85), los modelos entrenados con el conjunto híbrido lograron una mejora marginal en su capacidad discriminativa. Esta ganancia de rendimiento probablemente se explica por la incorporación de variables que reflejan estados acumulados posteriores al evento, lo cual, si bien puede ser útil para comprender correlaciones retrospectivas, compromete su aplicabilidad prospectiva.

Figura 10. Curvas ROC comparativas de modelos de machine learning para la predicción de vinculación laboral utilizando variables censuradas temporalmente y no censuradas.



Curvas ROC (Receiver Operating Characteristic) de los modelos de machine learning entrenados para predecir la probabilidad de vinculación laboral simultánea con la academia y el sector público o privado. El eje X representa la tasa de falsos positivos (1 - especificidad), y el eje Y representa la tasa de verdaderos positivos (sensibilidad). La línea diagonal discontinua corresponde a un modelo sin capacidad discriminativa (AUC = 0.50).

3.5.2. Métricas de rendimiento del modelo

Se evaluó el desempeño de múltiples algoritmos de ML entrenados con un conjunto extendido de variables predictoras, que incluyó tanto variables censuradas temporalmente como aquellas sin trazabilidad de fecha. Este enfoque híbrido se consideró como un modelo exploratorio complementario, útil para estimar el impacto adicional de variables acumuladas al cierre del estudio, aunque no aplicable directamente en contextos de predicción prospectiva.

Los resultados se muestran en la Tabla 3. El modelo Gradient Boosting alcanzó una exactitud (accuracy) de 0.90, precisión de 0.92, sensibilidad de 0.93, especificidad de 0.84, F1-score de 0.90 y un AUC-ROC de 0.89, representando uno de los dos algoritmos de mayor rendimiento global. De forma equivalente, el Random Forest mostró valores idénticos de accuracy (0.90) y sensibilidad (0.93), con una ligera ventaja en precisión (0.93) y especificidad (0.85), y el mismo F1-score (0.90) y AUC (0.89) que Gradient Boosting.

El modelo XGBoost también presentó un alto rendimiento, con accuracy = 0.88, precisión = 0.94, sensibilidad = 0.89, F1-score = 0.89 y AUC = 0.88. Por su parte, la Red Neuronal Multicapa (MLP Classifier) alcanzó un AUC de 0.88, F1-score de 0.89 y sensibilidad de 0.91, confirmando su solidez como clasificador no lineal. En contraste, el modelo SVC obtuvo la mayor sensibilidad (0.95), pero con especificidad reducida (0.78), lo que afectó el balance general a pesar de un AUC aceptable (0.86).

En conjunto, todos los modelos del enfoque exploratorio híbrido superaron las métricas observadas en sus equivalentes entrenados con variables censuradas exclusivamente (modelo principal). La mejora más notoria se observó en la precisión y el F1-score, particularmente en los modelos Gradient Boosting y Random Forest (F1 = 0.90 vs. 0.87 en el modelo principal), lo cual refleja el incremento de información predictiva al incluir variables sin restricción temporal. Sin embargo, dado que estas variables pudieron haberse generado después del evento de interés, este modelo

híbrido debe interpretarse con fines descriptivos, exploratorios, y no como una herramienta aplicable a predicciones prospectivas.

Tabla 3. Comparación de métricas de rendimiento de modelos para la predicción de vinculación laboral utilizando variables censuradas temporalmente y no censuradas.

Modelo	Exactitud	PPV	Especificidad	Sensibilidad	F1-score	AUC	FPV
Decision Tree Classifier	0.83	0.88	0.78	0.85	0.83	0.82	0.73
Random Forest Classifier	0.9	0.93	0.85	0.93	0.9	0.89	0.85
Gradient Boosting Classifier	0.9	0.92	0.84	0.93	0.9	0.89	0.85
XGB Classifier	0.88	0.94	0.88	0.89	0.89	0.88	0.8
SVC	0.89	0.89	0.78	0.95	0.89	0.86	0.88
MLP Classifier	0.89	0.93	0.85	0.91	0.89	0.88	0.83

PPV: valor predictivo positivo; **AUC:** área bajo la curva ROC; **FPV:** valor predictivo negativo; **F1-score:** medida armónica entre precisión y sensibilidad; **SVC:** *Support Vector Classifier*; **MLP:** red neuronal multicapa.

Las métricas fueron calculadas sobre el conjunto de validación, utilizando únicamente variables predictoras con registro de fecha, las cuales fueron censuradas temporalmente hasta el momento de ocurrencia del desenlace.

IV. DISCUSIÓN

El presente estudio evidenció que los investigadores peruanos registrados en RENACYT en el área de Ciencias Médicas y de la Salud, con filiación académica, presentaron una mediana del índice H de 2 (RIQ: 0–4), reflejando bajo impacto bibliométrico. Además, el 66.9 % reportó vinculación laboral adicional con el sector público o privado. En este contexto, se desarrollaron y validaron modelos de ML para predecir dicha doble vinculación laboral. El modelo principal, construido con Gradient Boosting y entrenado exclusivamente con variables censuradas hasta la fecha del evento de interés, mostró un desempeño elevado (AUC: 0.85; exactitud: 0.87; sensibilidad: 0.91; precisión: 0.90). Las principales variables incluyeron producción científica, asesoría de tesis y proyectos financiados, aunque los valores SHAP revelaron efectos no uniformes, sugiriendo interacciones complejas en los perfiles académico-laborales.

En escenarios de interacción entre académicos de la universidad y el sector productivo, se ha documentado la existencia de vínculos contractuales simultáneos entre investigadores universitarios y empresas privadas. En universidades estadounidenses, aproximadamente el 4 % de los científicos reportaron empleo con empresas privadas mientras ocupaban cargos académicos, y un 18 % actuaban como consultores industriales remunerados (40). En el contexto europeo, un estudio representativo identificó que el 16 % de los investigadores presentaban una doble afiliación formal con organizaciones del sector público o privado (39), proporción menor a la observada en el presente estudio, donde el 66.9 % de los investigadores reportaron una vinculación laboral adicional fuera del ámbito estrictamente académico. Estas diferencias pueden deberse a las distintas áreas de poblaciones universitarias estudiadas, el presente estudio solo aborda investigadores del área de Ciencias Médicas y de la Salud.

Si bien algunos trabajos sobre investigación en gestión en América Latina han descrito el crecimiento de la producción científica en la región y los desafíos estructurales como la inestabilidad institucional y las limitaciones de las carreras académicas (41), no se han identificado estudios que presenten datos específicos ni porcentajes sobre la magnitud del fenómeno de vinculaciones laborales simultáneas entre investigadores y académicos. No obstante, este hallazgo plantea la oportunidad

sobre el desarrollo de lineamientos que orienten una modalidad de vinculación laboral, considerando implicancias en la productividad científica, los posibles conflictos de interés, la independencia académica, así como la transferencia de conocimiento y tecnología.

En el contexto peruano, la doble vinculación puede interpretarse como una estrategia de adaptación frente a un ecosistema de ciencia, tecnología e innovación (CTI) caracterizado por financiamiento e infraestructura limitados, débil promoción del capital humano especializado y escasos incentivos para el desarrollo sostenido de la CTI. Estas condiciones, reconocidas en la Política Nacional de CTI, dificultan el fortalecimiento de capacidades institucionales y la consolidación de trayectorias científicas a largo plazo (42). En este escenario, los vínculos laborales adicionales permitirían acceder a ingresos complementarios, mantener la práctica profesional o participar en investigación aplicada.

Por otro lado, la doble vinculación también puede responder a factores estructurales del sector salud. En nuestra muestra, el 27.6 % de los investigadores tenían formación médica y otro 20.5 % correspondía a profesiones con funciones asistenciales. En estos casos, la coexistencia de actividades académicas y sanitarias es habitual y responde a una lógica institucional que articula docencia, investigación y ejercicio profesional, como ha sido documentado en la literatura sobre educación en salud (43,44).

No obstante, aunque la afiliación a instituciones puede facilitar el acceso a recursos complementarios, redes colaborativas y oportunidades de financiamiento, este esquema no garantiza mayor productividad científica si no está acompañado de condiciones estructurales favorables. Evidencias previas muestran que, sin tiempo protegido ni entornos institucionales que valoren la investigación, el potencial de estas configuraciones laborales puede verse limitado (45). Esto podría contribuir a explicar la baja mediana del índice H observada en nuestra muestra, incluso entre quienes presentaban doble vinculación.

En investigaciones previas se ha reportado que la participación en redes de colaboración científica, particularmente en contextos con recursos limitados, se

asocia con mayor productividad en publicaciones (46). Nuestros hallazgos evidencian que un incremento en la producción científica aparece después de la afiliación laboral en sectores públicos o privados, sin embargo, esto fue observado solo en un grupo reducido de investigadores del RENACYT en el área de Ciencias Médicas y de la Salud.

Por otro lado, Hottenrott et al. observaron que autores con tener adicionalmente otras afiliaciones suelen estar presentes en revistas de alto impacto y que la afiliación dual se incrementa con la diversificación temática y la visibilidad de la investigación (47). Mientras que, en otros entornos autores con múltiples afiliaciones actuaron como “puentes” entre organizaciones académicas y productivas, impulsados por proximidad institucional y científica (48). Estos hallazgos respaldan los casos donde investigadores con trayectorias balanceadas pueden servir de enlace efectivo entre diversos entornos laborales y de investigación.

Nuestros hallazgos evidencian que los indicadores de producción científica, como publicaciones indexadas en Scopus, proyectos financiados y tesis dirigidas, no presentan una relación uniforme con la probabilidad de una mayor vinculación laboral con el sector público o privado. El modelo de mejor desempeño fue Gradient Boosting, aplicado como algoritmo supervisado de ML en un enfoque de inteligencia artificial explicable, utilizando variables con censura temporal para predecir vinculación laboral. En concordancia con la literatura, las técnicas de ML han sido empleadas en la gestión del talento para predecir deserción, adecuación ocupacional y satisfacción profesional (8,12,13). En Perú, se han utilizado redes neuronales para estimar la satisfacción laboral docente (18) y se han comparado algoritmos en pruebas de aptitud ocupacional (19). No obstante, aunque existen aproximaciones exploratorias mediante minería de texto sobre currículums de investigadores peruanos (20), la evidencia empírica basada en modelos predictivos sobre su vinculación laboral continúa siendo limitada.

En nuestro estudio se evidencio que la productividad científica, por sí sola, no predice la probabilidad de una vinculación adicional a la académica. Esto se observa, por ejemplo, en investigadores orientados a la ciencia básica que, a pesar de una alta productividad, incluyendo publicaciones mayoritarias en revistas Q1, presentan

menor probabilidad de vincularse a sectores externos (47,49). En el caso peruano, entre los investigadores del área de Ciencias Médicas y de la Salud, se identificó que las publicaciones en revistas Q1 y Q2 contribuyeron de manera mixta o incluso negativa a la probabilidad de doble afiliación. Una posible explicación es que la producción de alto impacto, centrada en temáticas básicas, no necesariamente responde a las demandas prácticas del sector salud público-privado, limitando así su conexión con estos ámbitos.

Los valores SHAP en el estudio revelaron que los predictores en el modelo de ML tuvieron efectos no lineales e interacciones complejas entre variables. Esta interpretación va en línea con la literatura sobre modelos explicables en inteligencia artificial, donde se destaca la utilidad de SHAP para descomponer contribuciones individuales y capturar patrones no evidentes en modelos clásicos (50). Lundberg et al. han demostrado que los métodos basados en valores de Shapley ofrecen ventajas clave al atribuir la importancia de cada predictor en contextos con alta heterogeneidad, como ocurre en perfiles académicos con trayectorias múltiples (51).

Desde una perspectiva metodológica y de transparencia, la aplicación de valores SHAP en nuestro estudio representa una buena práctica alineada con estándares emergentes en la literatura de ciencias de la salud. En contextos clínicos, esta técnica se utiliza para interpretar modelos predictivos a nivel individual, permitiendo comprender cómo cada variable influye en la probabilidad de un desenlace (52). Al trasladar este enfoque al análisis de perfiles de investigadores, es posible descomponer la contribución específica de cada predictor con un nivel de detalle comparable al observado en sistemas de apoyo clínico, mejorando así la interpretabilidad de los modelos en el ámbito académico-laboral.

El uso del modelo híbrido, que incorporó variables sin censura temporal, resultó en una mejora marginal del desempeño predictivo (AUC de 0.85 a 0.89), en concordancia con estudios que reportan incrementos similares al incluir variables acumuladas a lo largo del tiempo (53). No obstante, el uso de información posterior al evento introduce un sesgo retrospectivo que compromete la validez del modelo para aplicaciones prospectivas, al generar estimaciones optimistas difíciles de reproducir en contextos reales. Este hallazgo resalta la solidez del enfoque basado en

censura temporal adoptado en el presente estudio, el cual preserva la integridad metodológica y contribuye a la aplicabilidad en escenarios de toma de decisiones futuras.

Los hallazgos de este estudio ofrecen implicancias relevantes tanto para la formulación de políticas de investigación como para la gestión estratégica del talento científico en países de renta media como Perú. Al aplicar modelos de ML explicables para predecir la vinculación laboral en investigadores del área de Ciencias Médicas y de la Salud, este trabajo introduce una metodología innovadora que trasciende los enfoques descriptivos tradicionales y permite identificar patrones complejos de comportamiento académico-laboral.

Además, este estudio plantea nuevas líneas de trabajo en la intersección entre ciencia de datos, política científica y análisis de carrera investigadora. La aplicabilidad de los modelos predictivos desarrollados podría extenderse a otros grupos disciplinares, permitiendo explorar factores relacionados a migración científica, inserción en la industria o patrones de internacionalización, campos que ya han sido priorizados en los informes oficiales de Perspectivas de la Ciencia, la Tecnología y la Innovación de la OCDE 2023 (54).

Así mismo, el enfoque del presente estudio basado en inteligencia artificial explicable aplicado a trayectorias científicas representa una innovación metodológica de alto potencial, al brindar una herramienta analítica que puede integrarse en plataformas de monitoreo de talento científico, en concordancia con propuestas recientes sobre gobernanza inteligente en Ciencia, Tecnología e Innovación (55). Así, este estudio no solo contribuye al entendimiento de las dinámicas laborales de los investigadores en el área de salud, sino que abre posibilidades para la planificación prospectiva del capital humano en ciencia.

Por otro lado, este estudio presenta limitaciones que deben ser consideradas. En primer lugar, el diseño transversal y el uso de datos secundarios auto declarados en la plataforma CTI Vitae podrían introducir sesgos de información, debido a posibles omisiones o registros incompletos. En segundo lugar, la definición del desenlace principal se basó en las afiliaciones institucionales declaradas, sin información

detallada que podría limitar la precisión del modelo. Asimismo, aunque se aplicaron técnicas de machine learning explicables, los patrones observados pueden estar influenciados por interacciones complejas entre variables. Por último, los modelos fueron entrenados únicamente en investigadores del área de Ciencias Médicas y de la Salud, por lo que los resultados deben interpretarse con cautela y no generalizarse a otros campos sin validaciones adicionales.

V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Este estudio evidencio que una mayor productividad académica no se relaciona de manera consistente con una mayor probabilidad de vinculación laboral adicional en el sector público o privado. Por el contrario, se observaron efectos mixtos y dependientes del contexto profesional del investigador, lo que sugiere que la doble afiliación responde a múltiples factores, incluyendo aspectos estructurales del ecosistema de ciencia y tecnología en el Perú.

Este estudio desarrolló y validó modelos de ML para predecir la doble vinculación laboral en investigadores del área de Ciencias Médicas y de la Salud registrados en RENACYT con filiación académica. El modelo principal, basado en Gradient Boosting con censura temporal, mostró un rendimiento robusto y permitió identificar patrones complejos e interacciones no lineales entre variables de producción científica y académica.

El estudio aporta evidencia empírica sobre las trayectorias laborales de los investigadores en salud y destacan el valor de los enfoques explicables de inteligencia artificial para analizar fenómenos complejos en sistemas de investigación nacionales.

5.2. Recomendaciones

Para CONCYTEC y gestores del sistema nacional de CTI:

Se recomienda mejorar la calidad, trazabilidad y actualización de los registros disponibles en plataformas como CTI Vitae, a fin de potenciar su utilidad como herramienta para el análisis longitudinal de trayectorias científicas y para la formulación de políticas basadas en evidencia. Asimismo, que la plataforma CTI Vitae cuente con un registro estandarizado y sistemático de la vinculación laboral, incorporando criterios que permitan explicitar con mayor claridad el tipo de función desempeñada y su contribución a la competitividad del sector privado o al bienestar público.

Esta evidencia empírica permitirá analizar con mayor profundidad el fenómeno de la doble afiliación en un país en desarrollo como el Perú. Sobre esta base, se sugiere elaborar lineamientos normativos que regulen dicha modalidad de vinculación, garantizando el cumplimiento de las funciones académicas esenciales, la integridad científica, y al mismo tiempo fomentando la transferencia de conocimiento, la

innovación aplicada y el fortalecimiento de servicios públicos sustentados en evidencia científica.

Para universidades:

Fortalecer los mecanismos institucionales orientados a promover trayectorias científicas estables, articuladas y sostenibles, considerando la naturaleza multidimensional del perfil de los investigadores. Asimismo, fomentar la interoperabilidad y la colaboración efectiva con entidades del sector público o privado, reconociendo el valor estratégico de los perfiles híbridos en contextos académicos, científicos y sanitarios.

Para empleadores del sector salud e instituciones públicas y privadas:

Evaluar el uso potencial de la información consolidada en plataformas como CTI Vitae, complementada con herramientas de inteligencia artificial, para fortalecer los procesos de identificación, selección y planificación del talento humano en investigación. El análisis responsable apoyado en tecnologías de IA podría contribuir a una mejor articulación entre el desarrollo científico y las necesidades del sistema de salud, en coordinación con el ámbito universitario.

Para investigadores y académicos de otras áreas del conocimiento:

Se recomienda replicar este enfoque metodológico en otras áreas conforme a la clasificación OCDE, a fin de identificar si los patrones observados son específicos del campo de ciencias médicas y de la salud o reflejan fenómenos estructurales del ecosistema científico nacional. Las futuras investigaciones podrían incorporar variables adicionales vinculadas a redes de colaboración, motivaciones profesionales, condiciones laborales e incentivos institucionales.

Asimismo, se sugiere explorar en futuras investigaciones el impacto diferenciado de las modalidades de vinculación laboral en relación con la competitividad del sector privado o el bienestar público. Esta línea de investigación, mediante enfoques cualitativos o mixtos que integren minería de texto, entrevistas semiestructuradas y verificación institucional, permitiría abordar una tipología más precisa y funcional, superando las limitaciones de los registros administrativos disponibles en plataformas como CTI Vitae.

VI. REFERENCIAS BIBLIOGRÁFICAS

1. Kaiser U, Kongsted HC, Laursen K, Ejlsing AK. Experience matters: The role of academic scientist mobility for industrial innovation. *Strategic Management Journal*. 2018; 39 (7): 1935-1958. Doi: <https://doi.org/10.1002/smj.2907>
2. D'Este P, Perkmann M. Why do academics engage with industry? The entrepreneurial university and individual motivations. *J Technol Transf*. 2011; 36: 316-339. Doi: <https://doi.org/10.1007/s10961-010-9153-z>
3. Balsmeier, B., Pellens, M. How much does it cost to be a scientist?. *J Technol Transf*. 2016; 41: 469-505. Doi: <https://doi.org/10.1007/s10961-014-9388-1>
4. Van Zyl-Cillie MM, Bührmann JH, Blignaut AJ, Demirtas D, Coetzee SK. A machine learning model to predict the risk factors causing feelings of burnout and emotional exhaustion amongst nursing staff in South Africa. *BMC Health Serv Res*. 2024 Dec 31;24(1):1665. doi: 10.1186/s12913-024-12184-5.
5. Leist AK, Klee M, Kim JH, Rehkopf DH, Bordas SPA, Muniz-Terrera G, Wade S. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Sci Adv*. 2022 Oct 21;8(42): eabk1942. Doi: 10.1126/sciadv.abk1942.
6. Hannaford L, Cheng X, Kunes-Connell M. Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. *Nurse Educ Today*. 2021; 99:104784. Doi: 10.1016/j.nedt.2021.104784.
7. Broda MD, Bogenschutz M, Dinora P, Prohn SM, Lineberry S, Ross E. Using Machine Learning to Predict Patterns of Employment and Day Program Participation. *Am J Intellect Dev Disabil*. 2021; 126(6): 477-491. Doi: 10.1352/1944-7558-126.6.477.
8. Wu W, Fukui S. Using Human Resources Data to Predict Turnover of Community Mental Health Employees: Prediction and Interpretation of Machine Learning Methods. *Int J Ment Health Nurs*. 2024; 33(6): 2180-2192. Doi: 10.1111/inm.13387.
9. Fukui S, Wu W, Greenfield J, Salyers MP, Morse G, Garabrant J, et al. Machine Learning with Human Resources Data: Predicting Turnover among Community Mental Health Center Employees. *J Ment Health Policy Econ*. 202; 26(2):63-76. PMID: 37357871

10. García-Peñalvo F, Cruz-Benito J, Martín-González M, Vázquez-Ingelmo A, Sánchez-Prieto JC, Therón R. Proposing a Machine Learning Approach to Analyze and Predict Employment and its Factors. *Int J Interact Multimed Artif Intell.* 2018; 5(2): 19–26. doi:10.9781/ijimai.2018.02.002.
11. Clemente CJ. Utilizing machine learning techniques in predicting job viability of information technology program graduates [doctoral dissertation]. Macon (GA): Middle Georgia State University; 2023.
12. Navarro Mazo, Nogueira Júnior MP, Soares de Quadros AR, Vieira A, Brandão WC. Predicting employee turnover using personality assessment: a data-driven approach. In: *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025).* 2025; 1: 582–589. Doi:10.5220/0013436400003929.
13. Wingrove P, Liaw W, Weiss J, Petterson S, Maier J, Bazemore A. Using Machine Learning to Predict Primary Care and Advance Workforce Research. *Ann Fam Med.* 2020; 18(4): 334-340. Doi: 10.1370/afm.2550.
14. Chen HH, Lu HH, Weng WH, Lin YH. Developing a Machine Learning Algorithm to Predict the Probability of Medical Staff Work Mode Using Human-Smartphone Interaction Patterns: Algorithm Development and Validation Study. *J Med Internet Res.* 2023; 25: e48834. Doi: 10.2196/48834.
15. Gupta A, Chadha A, Tiwari V, Varma A, Pereira V. Sustainable training practices: predicting job satisfaction and employee behavior using machine learning techniques. *Asian Bus Manag.* 2023: 1-24. Doi: 10.1057/s41291-023-00234-5.
16. Thomas D, Li Y, Ukwuoma CC, Dossa J. Explainability of artificial neural network in predicting career fulfilment among medical doctors in developing nations: Applicability and implications. *Soc Sci Med.* 2024; 360: 117329. Doi: 10.1016/j.socscimed.2024.117329.
17. Habehh H, Gohel S. Machine Learning in Healthcare. *Curr Genomics.* 2021; 22(4): 291-300. Doi: 10.2174/1389202922666210705124359.
18. Espinosa-Pinos, C.A.; Acosta-Pérez, P.B.; Valarezo-Calero, C.A. Applying Classification Techniques in Machine Learning to Predict Job Satisfaction of University Professors: A Sociodemographic and Occupational Perspective. 2024; 13(12): 344. Doi: <https://doi.org/10.3390/computers13120344>
19. Charapaqui-Miranda S, Arapa-Apaza K, Meza-Rodríguez M, Chacon-Torrico H. Comparing predictive machine learning algorithms in fit for work occupational

- health assessments. In: Information Management and Big Data. Communications in Computer and Information Science, vol 1070. Lossio-Ventura JA, Condori-Fernandez N, Valverde-Rebaza JC, editores. Cham (Suiza): Springer International Publishing; 2020. p 218–225. Doi:10.1007/978-3-030-46140-9_21
20. Chire Saire JE, Apaza Alanoca H. Text Mining over Curriculum Vitae of Peruvian Professionals using Official Scientific Site DINA, International Computer Symposium (ICS). 2020; 105-109. Doi: 10.1109/ICS51289.2020.00030.
 21. Flessa S, Huebner C. Innovations in Health Care-A Conceptual Framework. *Int J Environ Res Public Health*. 2021; 18(19): 10026. Doi: 10.3390/ijerph181910026.
 22. Martínez C, Parlane S. Academic scientists in corporate R&D: A theoretical model. *Research Policy*. 2023; 52(5): 104744. Doi: <https://doi.org/10.1016/j.respol.2023.104744>
 23. Dolmans SAM, Walrave B, Read S, Van Stijn N. Knowledge transfer to industry: how academic researchers learn to become boundary spanners during academic engagement. *J Technol Transf*. 2022; 47: 1422–1450. Doi: <https://doi.org/10.1007/s10961-021-09882-1>
 24. Vyas JM, Rajagopal J, Sokol CL, Wein MN, Mansour MK, Corey KE, et al. The Great Opportunity: Cultivating Scientific Inquiry in Medical Residency. *J Infect Dis*. 2018; (suppl 1): S44-S48. Doi: 10.1093/infdis/jiy200.
 25. Siemens DR, Punnen S, Wong J, Kanji N. A survey on the attitudes towards research in medical school. *BMC Med Educ*. 2010; 10:4. Doi: 10.1186/1472-6920-10-4.
 26. Mäkinen EI, Sapir A. Making Sense of Science, University, and Industry: Sensemaking Narratives of Finnish and Israeli Scientists. *Minerva*. 2023;61(2):175-198. Doi: 10.1007/s11024-022-09485-6.
 27. Wang Q, Ma J, Liao X, Du W. A context-aware researcher recommendation system for university-industry collaboration on R&D projects. *Decision Support Systems*. 2017; 103: 46-57. Doi: <https://doi.org/10.1016/j.dss.2017.09.001>
 28. Duval-Couetil N, Ladisch M, Yi S. Addressing academic researcher priorities through science and technology entrepreneurship education. *J Technol Transf*. 2021; 46: 288-318. Doi: <https://doi.org/10.1007/s10961-020-09787-5>.
 29. King M. What makes a successful corporate investigator: An exploration of private investigators attributes. *Journal of Financial Crime*. 2020; 27(3): 701-714. Doi: <https://doi.org/10.1108/JFC-02-2020-0019>

30. Mikalef P, Giannakos MN, Pappas IO, Krogstie J. The human side of big data: Understanding the skills of the data scientist in education and industry. IEEE Global Engineering Education Conference (EDUCON). 2018; 503-512. Doi: 10.1109/EDUCON.2018.8363273.
31. Comisión Consultiva de Ciencia, Tecnología e Innovación (CCCTI). Primer Informe: Acciones para el potenciamiento de la ciencia, la tecnología y la innovación en el Perú. Lima-Perú, 2021. [Acceso 20 de enero 2024]. Disponible en:
https://portal.concytec.gob.pe/images/noticias/Comisi%C3%B3n_Consultiva_Primer_Informe_280621__versi%C3%B3n_final_1.pdf
32. Ministerio de la Producción. Las MIPYME en cifras 2022. [Acceso 22 de enero 2024]. Lima-Perú, 2023. Disponible en:
<https://ogeiee.produce.gob.pe/index.php/en/shortcode/oe-documentos-publicaciones/publicaciones-anuales/item/1170-las-mipyme-en-cifras-2022>
33. Ministerio de la Producción. Resultados de la Encuesta Nacional de Innovación en la Industria Manufacturera y Empresas de Servicios Intensivas en Conocimiento – ENIIMSEC 2018. Lima-Perú, 2019. [Acceso 25 de enero 2024]. Disponible en: <https://ogeiee.produce.gob.pe/index.php/en/shortcode/estadistica-oe/estadisticas-innovacion>
34. Perú. Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC). Reglamento del Registro Nacional Científico, Tecnológico y de Innovación Tecnológica (RENACYT). Lima: CONCYTEC; [actualizado 2022 Mar; citado 2025 abril 2]. Disponible en: <https://servicio-renacyt.concytec.gob.pe/normativas/reglamento/>
35. Ley N°30220, Ley Universitaria. Lima-Perú, 2019. [Acceso 15 de marzo 2024]. Disponible en: <https://diariooficial.elperuano.pe/pdf/0021/ley-universitaria-30220.pdf>
36. Perú. Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC). Política Nacional para el Desarrollo de la Ciencia, Tecnología e Innovación Tecnológica (CTI) 2016–2021. Lima: CONCYTEC; 2016. Disponible en: <https://portal.concytec.gob.pe/images/publicaciones/politicas/politica-nacional-cti.pdf>

37. Perú. Ministerio de Economía y Finanzas. Decreto Supremo N° 345-2018-EF: aprueba la Política Nacional de Competitividad y Productividad. El Peruano. 31 dic 2018 [Internet]. Lima-Perú. [Acceso 15 de marzo 2024]. Disponible en: <https://www.gob.pe/institucion/mef/normas-legales/235968-345-2018-ef>
38. Marquina M. Between the global and the local: The study of the academic profession from a Latin American perspective. *Higher Education Governance & Policy*. 2020; 1(1), 63-75.
39. Cattaneo M, Horta H, Meoli M. Dual appointments and research collaborations outside academia: evidence from the European academic population. *Stud High Educ*. 2019;44(11):2066–80. Doi:10.1080/03075079.2018.1492534
40. Boardman PC, Ponomariov BL. University researchers working with private companies. *Technovation*. 2009; 29(2): 142–153. Doi: 10.1016/j.technovation.2008.03.008.
41. Aguinis H, Villamor I, Lazzarini SG, Vassolo RS, Amorós JE, Allen DG. Conducting management research in Latin America: Why and what's in it for you? *J Manag*. 2020;46(5):615–36. Doi:10.1177/0149206320901581.
42. Perú. Presidencia del Consejo de Ministros. Decreto Supremo N.º 093 2025 PCM que aprueba la Política Nacional de Ciencia, Tecnología e Innovación al 2030 (POLCTI). Lima: PCM; 15 jul 2025. Disponible en: https://www.gob.pe/institucion/pcm/normas-legales/6967622-093-2025-pcm?utm_source
43. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, Fineberg H, Garcia P, Ke Y, Kelley P, Kistnasamy B, Meleis A, Naylor D, Pablos-Mendez A, Reddy S, Scrimshaw S, Sepulveda J, Serwadda D, Zurayk H. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet*. 2010 Dec 4;376(9756):1923-58. Doi: 10.1016/S0140-6736(10)61854-5.
44. Woollard RF. Caring for a common future: medical schools' social accountability. *Med Educ*. 2006 Apr;40(4):301-13. Doi: 10.1111/j.1365-2929.2006.02416.x.
45. Sibley KM, Khan M, Touchette AJ, Crockett LK, Driedger SM, Gainforth HL, et al. Characterizing Canadian funded partnered health research projects between 2011 and 2019: a retrospective analysis. *Health Res Policy Syst*. 2023 Sep 8;21(1):92. Doi: 10.1186/s12961-023-01046-x.

46. Ynalvez MA, Shrum WM. Professional networks, scientific collaboration, and publication productivity in resource-constrained research institutions in a developing country. *Res Policy*. 2011 Mar;40(2):204–16. Doi: 10.1016/j.respol.2010.10.004
47. Hottenrott H, Rose M, Lawson C. The Rise of Multiple Institutional Affiliations in Academia. *JASIST*. 2021; 72(8): 1039-1058. Doi: <https://doi.org/10.1002/asi.24472>
48. Yegros-Yegros A, Capponi G, Frenken K. A spatial-institutional analysis of researchers with multiple affiliations. *PLoS ONE*. 2021; 16(6): e0253462. Doi: <https://doi.org/10.1371/journal.pone.0253462>.
49. Jurowetzki R, Hain D, Mateos-Garcia J, Stathoulopoulos K. The Privatization of AI Research(-ers): Causes and Potential Consequences. *arXiv* 2021;2102.01648.
50. Salih A, Raisi-Estabragh Z, Boscolo Galazzo I, Radeva P, Petersen SE, Lekadir K, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst*. 2024; 6(7): 2400304. Doi: <https://doi.org/10.1002/aisy.202400304>
51. Lundberg SM, Lee SI. A unified approach to interpreting model predictions [Preprint]. *arXiv*. 2017; arXiv:1705.07874. Doi: <https://doi.org/10.48550/arXiv.1705.07874>
52. Seyedamir Shobeiri. Enhancing Transparency in Healthcare Machine Learning Models Using Shap and Deeplift a Methodological Approach. *Iraqi Journal of ICT*. 2024; 7(2): 56-72. Doi: <https://doi.org/10.31987/ijict.7.2.285>.
53. Ghorbani A, Zou J. Data Shapley: equitable valuation of data for machine learning. In: Chaudhuri K, Salakhutdinov R, editors. *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*. 2019; 97:2242–51. Available from: <https://proceedings.mlr.press/v97/ghorbani19c.html>
54. OECD. *OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption*. Paris: OECD Publishing; 2023. Available from: <https://doi.org/10.1787/0b55736e-en>.
55. Cockburn IM, Henderson R, Stern S. The impact of artificial intelligence on innovation. NBER Working Paper No. 24449. Cambridge (MA): National Bureau of Economic Research. 2018. Doi: 10.3386/w24449