Modelo Predictivo de Campañas de Salud para Empresas Privadas del Perú Utilizando Aprendizaje Automatizado y OSEMN

Predictive Modeling of Health Campaigns for Private Sector Companies in Perú Using the Automated Learning and OSEMN Method

Nombres de Autores de Primera Institución
Línea 1 (Institución) Dep., Universidad, Organización
Línea 2 (Institución)
Línea 3: Ciudad, País
Línea 4: Correo electrónico

Nombres de Autores de Primera Institución
Línea 1 (Institución) Dep., Universidad, Organización
Línea 2 (Institución)
Línea 3: Ciudad, País
Línea 4: Correo electrónico

Nombres de Autores de Primera Institución
Línea 1 (Institución) Dep., Universidad, Organización
Línea 2 (Institución)
Línea 3: Ciudad, País
Línea 4: Correo electrónico

Resumen — **Debido a la nueva normalidad que nos dejó la pandemia, las empresas privadas han tenido que reforzar y mejorar las campañas de salud implementadas en favor del bienestar de sus colaboradores. Sin embargo, no se tiene justificación suficiente sobre cuáles de estas campañas de salud son las adecuadas a implementar ante la necesidad dentro de las empresas. Por ello, se propone un modelo predictivo de campañas de salud para una empresa privada de Lima - Perú utilizando 3 algoritmos de machine learning y la metodología OSEMN. La metodología consta de 5 fases: (1) obtención de datos, (2) depuración de datos, (3) exploración de datos, (4) modelamiento de algoritmos y (5) interpretación de resultados. El dataset utilizado proviene de 635 exámenes médicos de colaboradores de la empresa, identificando 14 variables. Los resultados evidenciaron que el modelo propuesto obtuvo un AUC de 0.998 superando el porcentaje mínimo óptimo de 0.90.**

Palabras clave — Campañas de salud; aprendizaje automatizado; salud de los empleados; análisis predictivo; gestión de la salud; bosques aleatorios; redes neuronales; regresión logística.

Abstract — **The new normal after pandemic has led the private companies to reinforce and improve their health campaigns for the well-being of their employees. However, an agreement has not yet been reached on what campaign is most suitable to implement in response to the need within companies. For this reason, a predictive model of health campaigns is proposed to a private company in Lima- Peru using three types of machine learning algorithms and the OSEMN method. The method includes five stages of data process: (1) Obtain, (2) Scrub, (3) Explore, (4) Model and (5) Interpret. The dataset used comes from 635 medical examinations carried out to the employees of the company where 14 variables were identified. The outcome shows that the model obtained an AUC of 0.998 surpassing the minimum optimal percentage of 0.90.**

Keywords — healthy campaign; machine learning; employee health; predictive analytics; health management; random forest; neural networks; logistic regression.

I. INTRODUCCIÓN

La pandemia, sin lugar a duda, ha evidenciado, en muchos países, la madurez en el sector salud para dar respuesta a una gran cantidad de emergencias y atenciones [1][2]. La salud en general comenzó a tener relevancia, pero como siempre suele darse, las atenciones y preocupaciones se dan de manera reactiva [2].

Lima, a finales del año 2022, mantiene una cantidad de 186,008 empresas en el sector privado lo que representa un 53.49% del total de las empresas privadas frente a otras regiones del país [3]. Se estima que la población económicamente activa (PEA) ocupada asciende a 5 millones 223 mil 400 personas, con un crecimiento de 3.1% respecto al año anterior [4]. Teniendo el análisis del campo general de impacto, se conoce también que a finales del año 2021, se estimó 16,925 millones de soles en gastos destinados al sector privado de salud y 31,104 millones de soles en gastos destinados al sector público [5]. Sin embargo, se evidencia que, para finales de este año, el 43.5% de la población reportó padecer alguna enfermedad crónica o problemas de salud en las 4 últimas semanas [6]. Con ello, se puede concluir que las campañas de salud o la inversión en este sector no está generando un impacto esperado.

Para mitigar este problema, han surgido diversos estudios que proponen la recopilación, análisis y predicción de la condición de salud de los colaboradores de una empresa. Algunos de ellos, proponen el uso del modelo predictivo de regresión, que utiliza técnicas de factorización de matriz neuronal extendida, procesamiento de lenguaje natural y un modelo basado en aprendizaje profundo [7][8] para el seguimiento de campañas de salud. Además, existen otros estudios [7][9] que complementan el estudio de la salud de los colaboradores con componentes cualitativos tales como el estrés y factores externos. También, existe un estudio [10] que propone investigar los beneficios nutricionales para ser transmitidos adecuadamente para tomar conciencia. Sin embargo, estos estudios no guardan relación con la toma de

acción preventiva frente a esta futura predicción crítica de salud.

Por tal motivo, el presente estudio propone un sistema para la predicción de campañas de salud en empresas privadas de Perú utilizando aprendizaje automatizado, con la finalidad de predecir las campañas de salud que se van a implementar en las empresas. La propuesta se realiza en cinco fases: (1) obtención de datos, (2) depuración de datos, (3) exploración de datos, (4) modelamiento de datos y (5) interpretación de resultados.

Este estudio presenta la siguiente estructura: Primero, la presentación de los proyectos relacionados en la sección 2. A continuación, los detalles de la ejecución de las campañas de salud en la sección 3. En la sección 4, se describe la propuesta y su ejecución. Finalmente, en la sección 5, los resultados obtenidos en el proceso de validación. Por último, se presentan las conclusiones y futuros proyectos en la sección 6.

II. TRABAJOS RELACIONADOS

La metodología implementada se basó en Wong et al. Proyecto. [11], en el cual se aplicaron las siguientes fases: (1) planificación, (2) desarrollo, (3) resultados y análisis. Para la etapa de planificación, se plantearon tres preguntas clave. La primera pregunta identifica los factores involucrados en campañas de salud, ¿Qué componentes encontramos en los modelos predictivos para campañas de salud? (P1). La segunda pregunta valida las tecnologías propuestas en los estudios, ¿Qué tecnologías se pueden usar para construir un sistema integrado de información para campañas de salud? (P2). La última pregunta se enfoca en las técnicas utilizadas por las diferentes soluciones para implementar y validar las campañas de salud, ¿Qué modelos predictivos existen para campañas de salud y cómo se validan? (P3).

Asimismo, antes de iniciar la fase de desarrollo, se realizaron filtros para obtener los artículos más relevantes relacionados con el tema propuesto. En Primer lugar, se definió el banco de journal bajo los siguientes criterios: innovación de sistemas aplicados, informes científicos, medicina ocupacional y ambiental, actividad física y salud. En segundo lugar, se definió que los artículos seleccionados deberían encontrarse en los metabuscadores IEEE Xplore, Scopus y EBSCO Host. En tercer lugar, los artículos relacionados deberían haber sido publicados entre inicios del 2020 y primer trimestre del 2023, además se deberían considerar las palabras clave campañas de salud, machine learning, salud de los empleados y salud en el trabajo. Finalmente, se definió considerar también los criterios de inclusión y exclusión (ver Tabla I).

TABLA I. CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN

Criterios de inclusión	Criterios de exclusión
Debe permitir responder a las preguntas de investigación	Artículos orientados para campañas de salud en el sector público
Debe presentar validación	Artículos de modelos no predictivos
Debe proponer un modelo de aprendizaje automatizado basado en campañas de salud	Artículos con antigüedad mayor a 3 años

Para la fase de desarrollo, utilizando los criterios ya expuestos se seleccionan los artículos potencialmente útiles para la investigación. Estos artículos fueron analizados en base a tres categorías: factores, tecnologías y técnicas.

A. *Factores*

Para abordar los problemas de campañas de salud se han aplicado diferentes factores, tales como información de salud, nutrición, actividad física, estrés, desorden alimenticio, influenza, higiene ocupacional, seguridad e información personal. El factor más utilizado fue de "información de salud". En [14] registraron información personal de los trabajadores de una fábrica en Japón, en [7] recopilaron información de salud del personal y datos técnicos y económicos, en [10] registraron datos de medicina general de pacientes de la India y EEUU, en [17] seleccionaron datos de accidentes y enfermedades laborales, en [19] registró información de salud a corto y largo plazo, en [20] registraron datos de pacientes de Nueva York para alfabetización en salud, en [21] recopilaron datos para promoción de la salud y vigilancia médica, en [22] registraron datos de salud de mujeres embarazadas en Países Bajos y China, en [25] recopilan información clínica para educación en salud, en [26] registraron información de los trabajadores de problemas de salud. Respecto a "nutrición", en [9] recopilaron información nutricional de los trabajadores de una empresa en Brasil, en [18] registraron la información de la actividad física y nutricional de los trabajadores, en [23] recopilaron los hábitos alimenticios de 15 mil trabajadores de ELSA-Brasil. Referente a "actividad física", en [11] registraron resultados de programas de trabajo basados en actividad física en las organizaciones de España, en [16] registraron información personal y de ejercicios realizados de los trabajadores de una empresa en Colombia, en [18] recopilaron información de la propia actividad física de los colaboradores empleando un aplicativo de salud. Respecto al factor "estrés", en [12] reunieron 440 mil reseñas en línea referentes al impacto del estrés en el trabajador, en [24] registraron datos sobre ansiedad y depresión producto del estrés por las largas horas de trabajo en empresas de tecnología. En cuanto a "desorden alimenticio", en [18] se registraron las experiencias en el control de peso en la población de Nueva Zelanda. Con respecto a "influenza", en [13] reunieron los patrones de comportamiento en cuanto al virus en la pandemia de 112,136 personas en 175 países. Referente a "higiene ocupacional", en [15] registraron actividades básicas de higiene ocupacional de diferentes años de una aseguradora en Colombia. En cuanto a "seguridad", en [17] revisaron 12 base de datos con información de las actividades a tiempo completo realizadas por los trabajadores de construcción civil y siderurgia. Referente a "información personal", en [26] registraron información laboral y personal de trabajadores de 80 empresas en 16 sectores económicos (ver Tabla II).

TABLA II. FACTORES DE CAMPAÑAS DE SALUD

Factores	Fuentes
Información de Salud	[14] [7] [10] [17] [19] [20] [21] [22] [25] [26]
Nutrición	[9] [18] [23]
Actividad física	[11] [16] [18]
Estrés	[12] [24]
Desorden alimenticio	[8]
Influenza	[13]
Higiene ocupacional	[15]
Seguridad	[17]
Información personal	[26]

B. Tecnologías

Se han encontrado diversas tecnologías para la implementación de campañas de salud, tales como de sistema [10], [15], [16], [9], [17], [19], [20], [22], [23], [24], [26], plataforma web [12], [13], [14], [21], aplicaciones [7], [25] y programas [18]. La tecnología más utilizada es la de "sistema" (ver Tabla III).

TABLA III. TECNOLOGÍAS EN CAMPAÑAS DE SALUD

Tecnología	Fuentes
Sistema	[10] [15] [16] [9] [17] [19] [20] [22] [23] [24] [26]
Plataforma web	[12] [13] [14] [21]
Aplicación	[7] [25]
Programa	[18]

C. Técnicas de IA

Diferentes técnicas de IA se han aplicado para abordar los problemas en las campañas de salud, tales como bosques aleatorios, redes neuronales, aprendizaje profundo, algoritmos de k-means, ensayos controlados aleatorios, arboles de decisión y regresión lineal multigrupo. La técnica de mayor uso fue de "bosques aleatorios". Referente a esta técnica, en [13] evaluaron el comportamiento de las personas en la pandemia en cada país, en [14] realizaron la predicción de las bajas por enfermedad en una fábrica japonesa, en [18] realizaron la predicción del porcentaje de encuestas de actividad física y salud, en [24] realizaron la predicción del porcentaje de estrés en los empleados del área de TI. Respecto a "redes neuronales", en [22] procesaron texto libre y lo clasificaron por categorías, en [23] realizaron recomendaciones dietéticas, en [26] Realizaron predicciones sobre la vida laboral y personal de los trabajadores. En cuanto a "regresión logística", en [30] Optimización de estrategias para identificación de pacientes con enfermedades cardiovasculares de alto riesgo, en [31] predicción de enfermedades cardiovasculares empleando algoritmos de aprendizaje automatizado. Referente a "aprendizaje profundo", en [12] evaluaron la productividad de los trabajadores, en [15] realizaron la reducción de riesgos de higiene ocupacional. Asimismo, respecto a "algoritmo k-means", en [7] realizaron la formación de grupos de información. Además, referente a "ensayos controlados aleatorios", en [16] evaluaron los efectos del ejercicio sobre el estrés. Por otro lado, respecto a "árboles de decisión", en [17] se utilizaron para prevenir accidentes laborales. Finalmente, referente a "regresión lineal multigrupo", en [19] compararon los efectos del liderazgo transformacional sobre el estrés (ver Tabla IV).

TABLA IV. TÉCNICAS DE IA APLICADAS EN EL SECTOR SALUD

Técnicas IA	Fuentes
Bosques aleatorios	[13] [14] [18] [24]
Redes neuronales	[22] [23] [26]
Regresión Logística	[30][31]
Aprendizaje profundo	[12] [15]
Algoritmo k-means	[7]
Ensayos controlados aleatorios	[16]
Árboles de decisión	[17]
Regresión lineal multigrupo	[19]

III. MODELO PROPUESTO

Esta sección explica el modelo para la implementación del sistema predictivo de campañas de salud para empresas privadas en Perú utilizando dos algoritmos de machine Learning: árboles de decisión y bosques aleatorios. Para ello se utilizará las fases de OSEMN [27].

A. Obtención de datos

Los conjuntos de datos iniciales brindados referentes al examen médico ocupacionales por la empresa en estudio contemplan 197 variables posible para ser evaluadas en el transcurso de 3 años entre el 2020 y 2022. La tabla muestra la cantidad de personal y registros obtenidos por cada año. Además, se tiene la información de los descansos médicos entre el 2020 y 2023. La información obtenida fue de unos 1059 registros en total (ver Tabla V).

TABLA V. CANTIDAD DE DATOS DE SALUD POR AÑO

Año	Cantidad personal	Cantidad registros
2020	349	68,753
2021	365	71,905
2022	429	84,513
Total	1,059	225,171

B. Depuración de datos

De los 197 atributos iniciales se seleccionaron 39 atributos principales los cuales fueron agrupados en 5 categorías. La Tabla VI muestra los atributos agrupados por categorías. Estas categorías fueron definidas por el médico ocupacional tomando como referencia la información histórica de salud de los trabajadores de la empresa obtenida años anteriores.

TABLA VI. ATRIBUTOS POR CATEGORÍAS

Categorías	Atributos
Información personal del paciente	Nombres (F01), DNI (F02), edad (F03), sexo (F04), fecha de nacimiento (F05).
Características laborales	Fecha de ingreso (F06), fecha de exámenes 2019 (F07), puesto (F08), años de trabajo (F09)
Información médica del paciente	Grupo sanguíneo (F10), habito nocivo al alcohol (F11), habito nocivo a la marihuana (F12), habito nocivo a la cocaína (F13), R.A.M u otra causa (F14), peso (F15), talla (F16), IMC (F17), FC (F18), FR (F19), PAD (F20), PAS (F21), EKG (F22).
Información de exámenes del paciente	Prueba de esfuerzo (F23), osteomuscular (F24), rayos x pulmonar (F25), espirometría conclusión (F26), audiometría ocupacional (F27), hemoglobina (F28), colesterol (F29), glucosa (F30), triglicéridos (F31), creatinina (F32), ácido úrico (F33), examen completo de orina (F34), psicología (F35).
Diagnóstico del paciente	Diagnóstico medicina (F36), diagnóstico oftalmología (F37), diagnóstico audiometría (F38), diagnóstico cardiología (F39).

De los 39 atributos obtenidos, para fines del presente estudio, se retiran los atributos cualitativos. Únicamente conservaremos los atributos que nos generen un valor cuantitativo medible que tenga relevancia para el entrenamiento. Esto también es definido por el médico ocupacional. Los registros se reducen a 14 atributos finales los cuales serán utilizados en el entrenamiento de los algoritmos

de clasificación. Estos atributos se agrupan en 3 categorías finales. La Tabla VII muestra los atributos finales agrupados por categoría.

TABLA VII. ATRIBUTOS POR CATEGORÍAS FINAL

Categorías	Atributos
Características laborales	Años de trabajo (F01)
Información médica del paciente	Habito nocivo al alcohol (F02), habito nocivo a la marihuana (F03), habito nocivo a la cocaína (F04), IMC (F05), FC (F06), FR (F07), PAD (F08), PAS (F09).
Información de exámenes del paciente	Hemoglobina (F10), colesterol (F11), glucosa (F12), triglicéridos (F13), creatinina (F14).

C. Exploración de datos

En esta fase, los 14 atributos encontrados en el paso anterior "depuración de datos de salud" se convierten en valores cuantificables y sirven como input de información en el entrenamiento de los algoritmos. Para ello emplearemos el concepto de correlación que nos indica si una variable es más o menos influyente sobre otra. En base a este concepto, se dejaron de lado los atributos de la categoría "información personal del paciente" debido a que no generan algún valor importante a considerar dentro de la correlación. En la Figura 1 se muestra la matriz de correlación entre las 14 variables finales seleccionadas. Los valores cuantificables que se otorgaron a cada variable dentro de la matriz fueron ingresados por el médico ocupacional, quien consideró el grado de influencia que tienen las variables entre sí. De esta forma se observa que el mayor valor de implicancia se da entre los atributos de hábitos nocivos a la marihuana (F03) y hábitos nocivos a la cocaína (F04). Además, en la Figura 2 se presenta el diagrama de barras, donde se muestran los atributos que tuvieron mayor relevancia, el atributo más sobresaliente fue el de "años de trabajo" (F01).

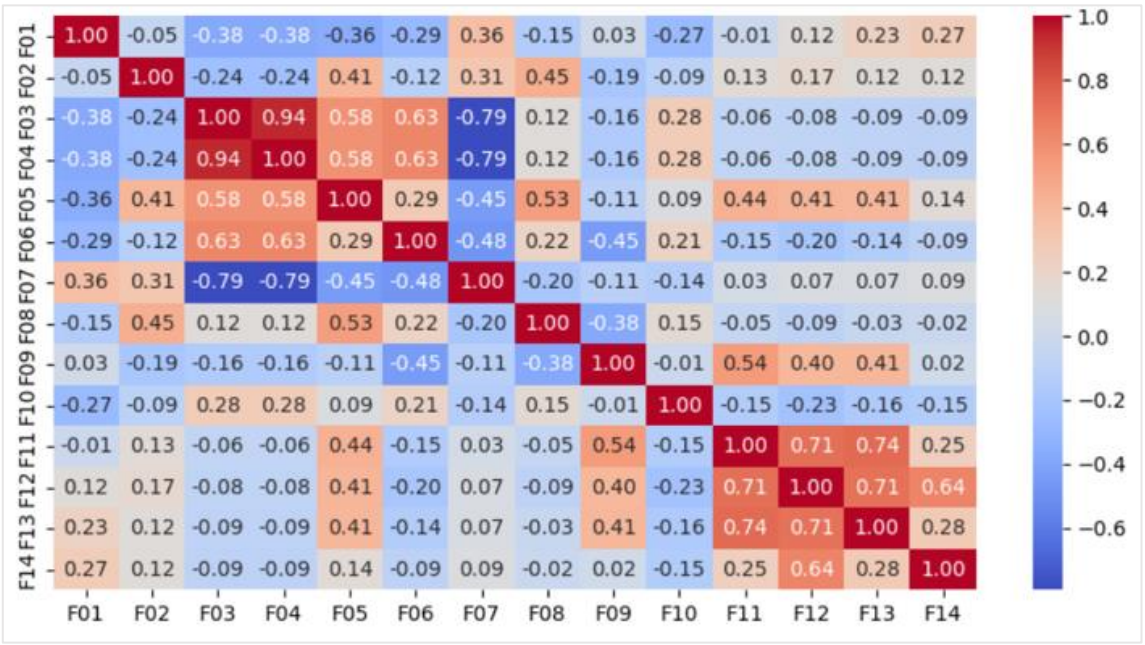

Figura 1. Matriz de correlación de atributos

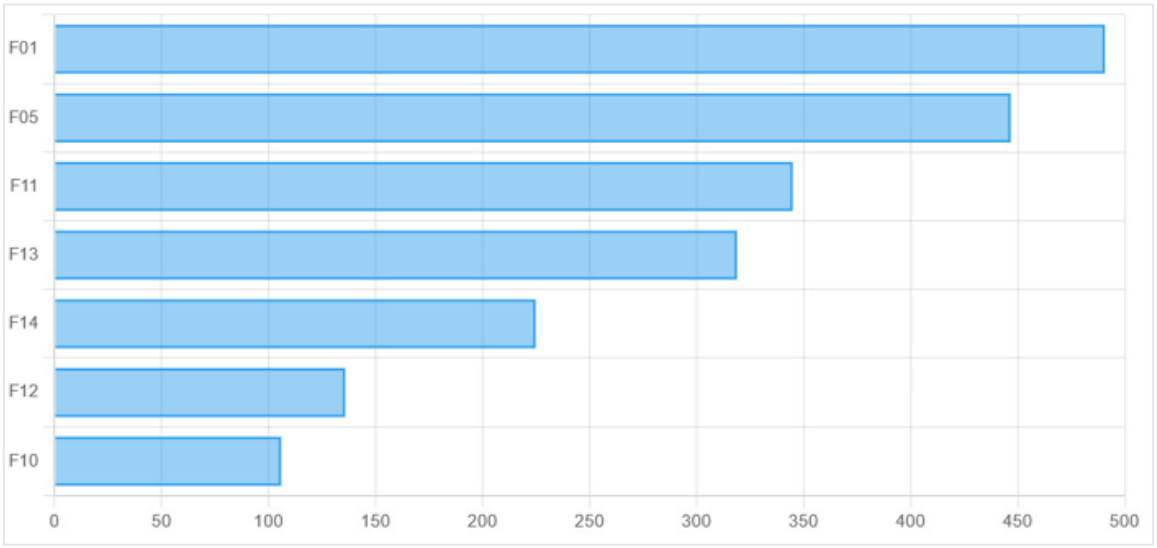

Figura 2. Cantidad de registros por atributos

D. Modelamiento de algoritmos

Existen diferentes algoritmos de clasificación que se pueden emplear para la predicción de campañas de salud. Considerando los algoritmos de decisión más utilizados en la literatura (ver Tabla IV), se han considerado los más utilizados: Bosques aleatorios (RF), redes neuronales (NN) y regresión logística (LR).

Bosques aleatorios (RF).

Algoritmo que se basa en el uso de múltiples arboles de decisión para la clasificación de los datos.

Redes neuronales (NN).

Algoritmo basado en neuronas y nodos artificiales. Los resultados se dividen en diferentes capas en base a los datos que procesa.

Regresión logística (LR).

Algoritmo de aprendizaje supervisado que se emplea en la clasificación. Se utiliza para predecir la probabilidad de que una observación pertenezca a dos o más clases.

Para comparar los resultados de los algoritmos en la predicción de campañas de salud se emplearán las métricas de precisión (1), exactitud (2), recuperación (3) y Puntaje F1 (4) respectivamente.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Se han empleado los siguientes términos:

- Verdadero positivo (TP): Predicción correcta sobre el uso de campañas de salud.
- Verdadero negativo (TN): Predicción correcta sobre las no uso de campañas de salud.
- Falso positivo (FP): Predicción incorrecta sobre el uso de campañas de salud.
- Falso Negativo (FN): Predicción incorrecta sobre el no uso de campañas de salud.

Se emplearon las siguientes métricas:

- Exactitud: Proporción de campañas de salud correctamente identificadas frente el número total de campañas.
- Precisión: Proporción de campañas de salud identificadas positivas frente al número de campañas de salud correctamente identificadas.
- Recuperación: Proporción de campañas de salud identificadas positivas capturadas.
- Puntaje F1: Rendimiento del modelo basado en la precisión y recuperación de campañas de salud.

Para el entrenamiento de los algoritmos de RF, NN y LR se utilizó la plataforma *Google Colab* y el lenguaje de programación *Python*. Para todos los casos se ha trabajado con el mismo data set inicial. La Figura 3 muestra el código en Python que se desarrolló para entrenar el modelo RF.

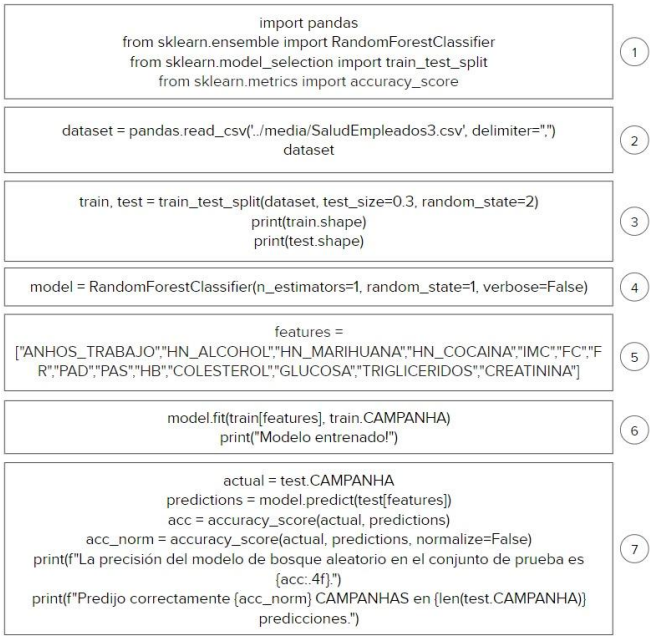

Figura 3. Pasos de entrenamiento de RF en Google Colab

Como primer paso, se importaron las librerías de *pandas* para proporcionar estructuras de datos de alto rendimiento y *sklearn* para el procesamiento de los datos. En el paso 2, cargamos el data set con la información que se utilizó en el entrenamiento del algoritmo RF. El data set contiene información de un total de 635 exámenes médicos realizados a los colaboradores de la empresa. La información se clasificó en trabajadores que no necesitan campañas de salud (NC), trabajadores que necesitan una campaña de salud (CA) y trabajadores que requieren una campaña de salud urgente (UC). En el paso 3, se configuraron los datos para el entrenamiento del modelo. Se empleó la proporción 70/30 para dividir la cantidad de registros, 444 registros para el entrenamiento y 191 registros para las pruebas. En el paso 4, se creó el modelo de RF donde se definió el número de estimadores, la aleatoriedad del algoritmo y la cantidad de información que se muestra durante su ejecución. En el paso 5, se definieron las características del dataset para el modelo. En el paso 6, se realizó el entrenamiento del modelo de RF utilizando como variable objetivo "campanha". Finalmente, se realizó el cálculo de la precisión del modelo en el conjunto de prueba en el paso 7.

IV. RESULTADOS Y DISCUSIÓN

E. Interpretación de resultados

Se emplearon distintas métricas de apoyo en la interpretación de los resultados de los algoritmos de RF, NN y LR. La Figura 4, Figura 5 y Figura 6 muestran la matriz de confusión de las predicciones realizadas de los algoritmos de RF, NN y LR. La información converge en la Tabla VIII donde se puede apreciar la comparación entre predicciones correctas e incorrectas de salud por cada algoritmo. La Figura 7 y Figura 8 muestran la comparación de las curvas de ROC de los algoritmos RF y LR respectivamente.

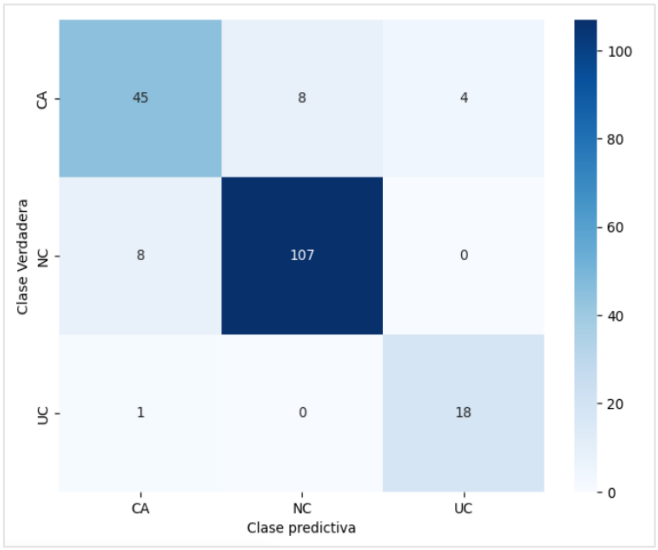

Figura 4. Matriz de confusión sobre las predicciones realizadas en RF

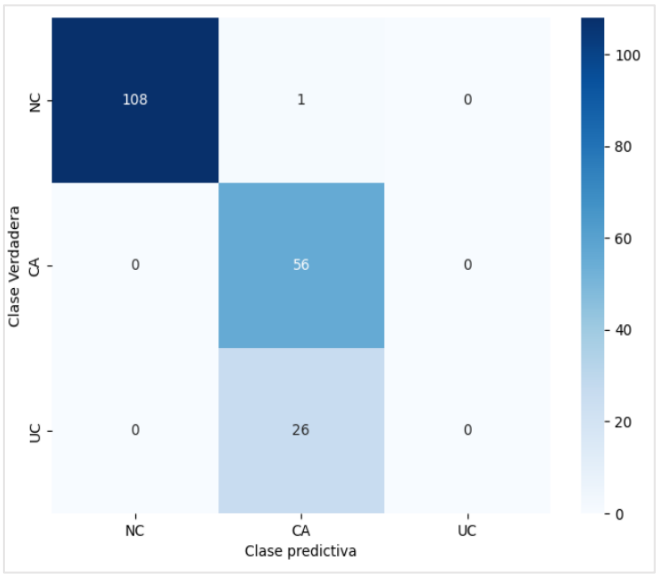

Figura 5. Matriz de confusión sobre las predicciones realizadas en NN

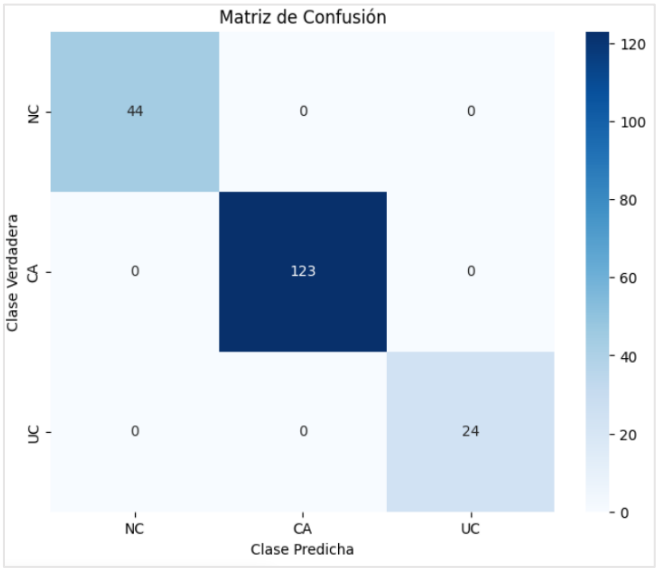

Figura 6. Matriz de confusión sobre las predicciones realizadas en LR

TABLA VIII. PREDICCIONES CORRECTAS E INCORRECTAS DE CAMPAÑAS DE SALUD POR ALGORITMO.

Algoritmo	Resultado	Predicciones correctas	Predicciones incorrecto
RF	CA	45	12
	NC	107	8
	UC	18	1
NN	NC	108	1
	CA	56	0
	UC	0	26
LR	NC	44	0
	CA	123	0
	UC	24	0

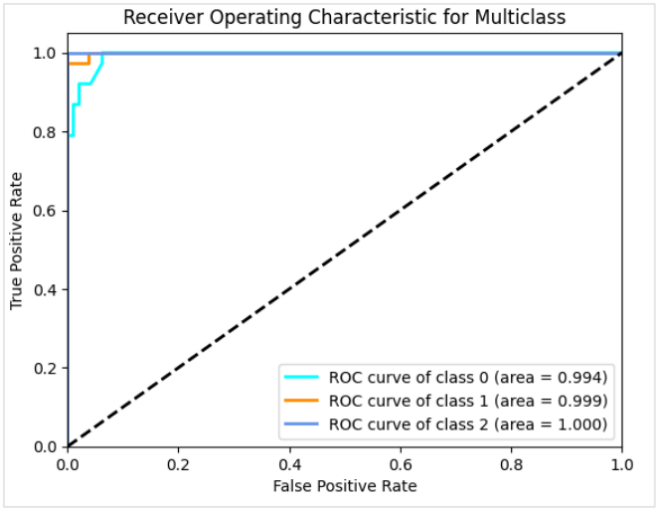

Figura 7. Comparación de curvas ROC algoritmo RF

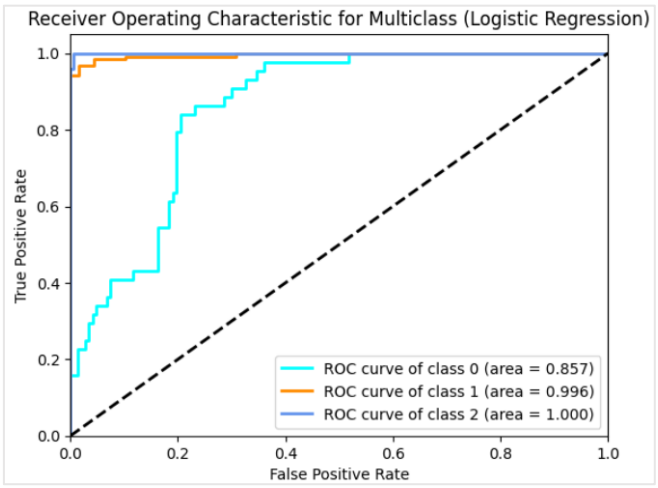

Figura 8. Comparación de curvas ROC algoritmo LR

La Tabla IV muestra la comparación de los resultados a las métricas establecidas anteriormente entre los algoritmos de RF, NN y LR. Observamos que el algoritmo de LR obtuvo el mayor porcentaje de exactitud con un 99.9%. En las métricas de recuperación observamos que el algoritmo de NN y LR obtuvieron mejores resultados que el algoritmo de RF. Podemos observar también que el algoritmo de NN y LR

tuvieron un mejor rendimiento que el algoritmo de RF, sin embargo, todos obtuvieron un rendimiento mayor al 90%.

TABLA IV. COMPARACIÓN DE MÉTRICAS POR ALGORITMO.

Algoritmo	Clasificación	Precisión	Recuperación	Puntaje F1	Exactitud	AUC
RF	CA	0.918	0.849	0.882	0.857	0.998
	NC	0.930	0.930	0.930	0.908	
	UC	1.000	0.947	0.973	0.947	
NN	NC	0.991	1.000	0.995	0.991	0.997
	CA	1.000	1.000	1.000	1.000	
	UC	1.000	1.000	1.000	1.000	
LR	NC	1.000	1.000	1.000	1.000	0.951
	CA	1.000	1.000	1.000	1.000	
	UC	1.000	1.000	1.000	1.000	

TABLA X. COMPARACIÓN CON OTROS ESTUDIOS.

Estudio	Objetivo	Clases	AUC	Cantidad de registros	Cantidad de variables
RF (enfoque propuesto)	Predicción de campañas de salud	3	0.998	635	14
NN (Enfoque propuesto)	Predicción de campañas de salud	3	0.997	635	14
LR (Enfoque propuesto)	Predicción de campañas de salud	3	0.951	635	14
RF [Aracena Et. al 2022] [28]	Predicción de estado de salud	2	0.98	300	5
RF [Thanki Et. al 2023] [29]	Predicción de estado de retina	2	0.988	650	9
NN [Thanki Et. al 2023] [29]	Predicción de estado de retina	2	0.819	650	9

F. Comparación con otros estudios

La Tabla X muestra la información de los modelos elaborados en el presente estudio y realiza una comparativa con similares investigaciones que emplean y clasifican también información de salud. Por ejemplo, podemos apreciar que, a nivel de efectividad, El modelo de RF de Aracena Et. al [28] muestra el AUC más alto (0.98) en comparación con el propuesto (0.998). Esto sugiere que el algoritmo entrenado es más efectivo en la predicción de resultados de salud. Es importante considerar otros factores, como la complejidad del modelo y el tamaño del conjunto de datos. Bajo esa premisa, el modelo propuesto utiliza un conjunto de datos más grande (635 registros) en comparación con Aracena Et. al [28] (300 registros). Un conjunto de datos más grande puede proporcionar más información para entrenar al modelo. Por otro lado, el modelo propuesto emplea más variables (14) en comparación con los modelos de Aracena Et. al [28] (5) y Thanki Et. al [29] (6). Esto nos indica que el enfoque propuesto considera más factores en su predicción. Es importante considerar que la elección del modelo debe basarse en un equilibrio entre precisión, complejidad y tamaño del conjunto de datos. Cada enfoque tiene sus ventajas y limitaciones, es importante considerar el contexto específico de la aplicación.

G. Construcción del sistema

El sistema se construyó con las herramientas de desarrollo en lenguaje C# .net, se emplearon las librerías chart.js para visualización de datos, html5 como estándar para la estructura y contenido de la página web, jQuery para la interacción entre la web y las aplicaciones desarrolladas, y Bootstrap para diseñar las plataformas. Se empleó el repositorio GitHub para el manejo de versiones del sistema. El desarrollo se hizo en n capas para la mejor administración de las dependencias. Por otro lado, se empleó el motor de base de datos SQL Server para almacenar, procesar y proteger los datos. Además, se utilizó un dominio registrado en GoDaddy con certificado de seguridad sobre un servidor de Windows sobre el cual se desplegó el sistema propuesto, tal como lo apreciamos en la Figura 9. El sistema se construyó a partir del modelo que tuvo mayor rendimiento el cual fue RF. En la Figura 10 podemos ver el diagrama C4 el cual modela la arquitectura software del sistema. El sistema muestra distintas gráficas que le permiten al médico ocupacional revisar el estado de salud general de los colaboradores. En la Figura 11 podemos apreciar la cantidad de colaboradores por años de trabajo, en la Figura 12 se muestra la cantidad de colabores por valor de salud y en la Figura 13 se aprecia la cantidad de colaboradores por tipo de campaña.

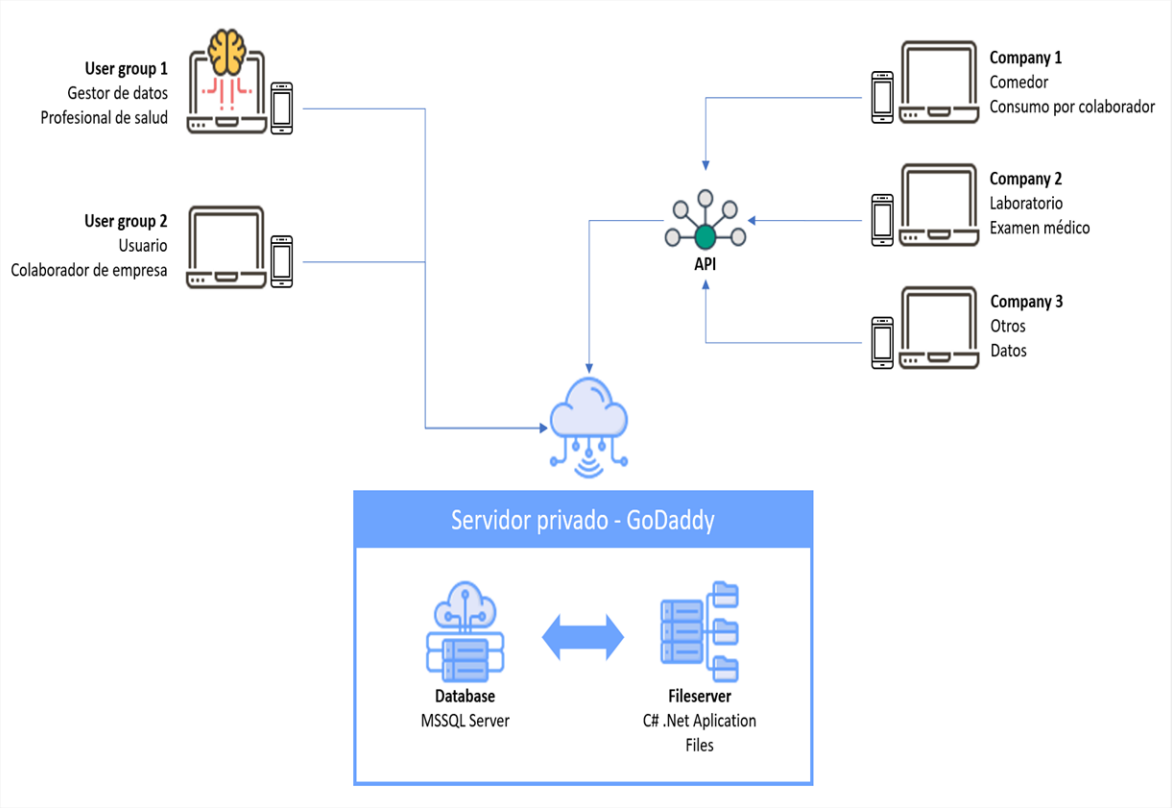

Figura 9. Diagrama general del sistema

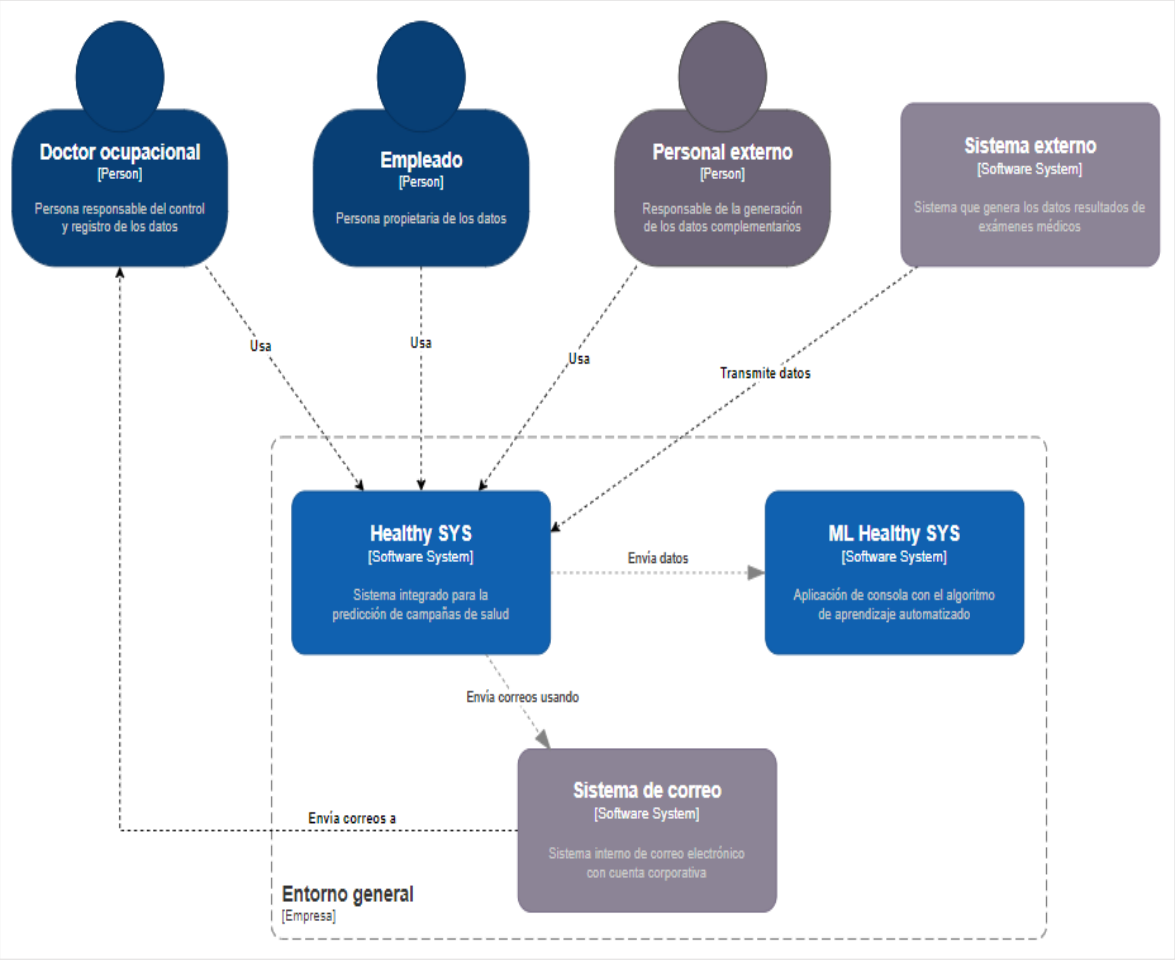

Figura 10. Diagrama C4 del sistema

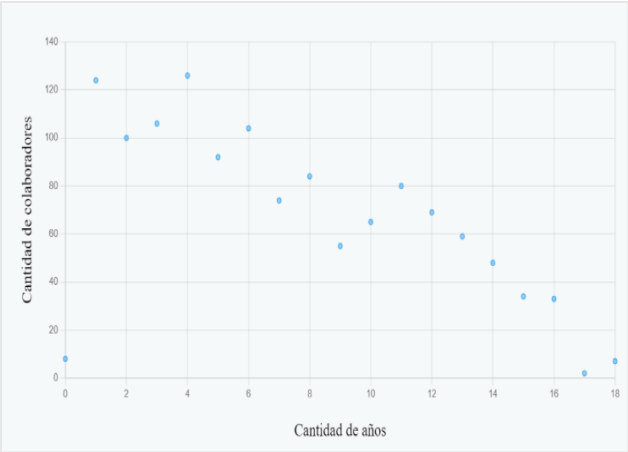

Figura 11. Cantidad de colaboradores por años de trabajo

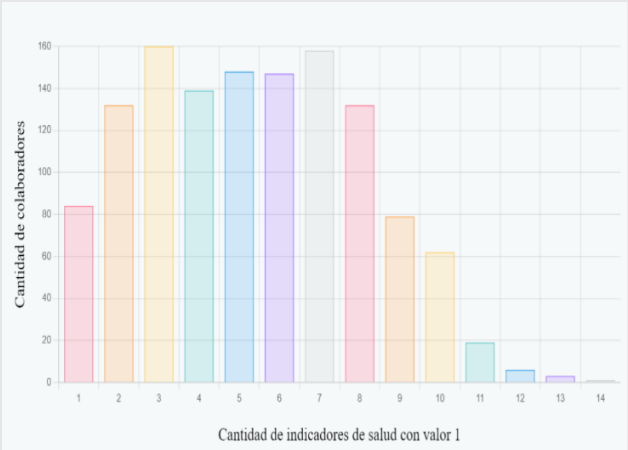

Figura 12. Cantidad de colaboradores por indicadores de salud

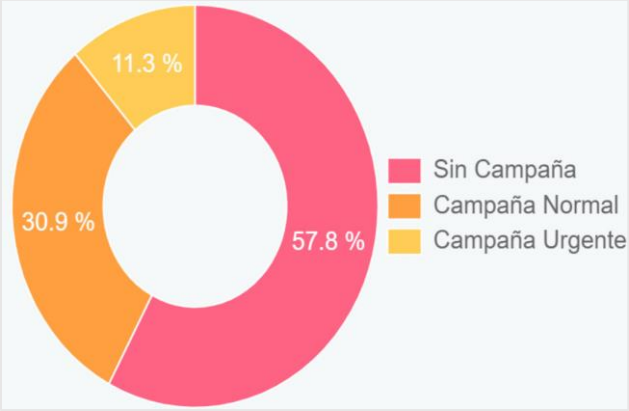

Figura 13. Cantidad de colaboradores por tipo de campaña

V. Conclusiones y Trabajos Futuros

En el presente trabajo se propuso un modelo predictivo de campañas de salud para empresas privadas del Perú utilizando aprendizaje automatizado y OSEMN. Para realizar las predicciones se empleó dos algoritmos de ML: Bosques aleatorios y redes neuronales. Siguiendo el marco de trabajo OSEMN, la propuesta se realizó en cinco fases: (1) obtención datos, (2) depuración datos, (3) exploración de datos, (4) modelamiento de algoritmos y (5) interpretación de resultados.

El dataset que se elaboró, se obtuvo a partir de la información histórica preparada por el medico ocupacional de la empresa en la que se centró el presente trabajo. Se obtuvieron 197 atributos iniciales y más de 223 mil registros de información de los principales datos de salud de los colaboradores.

En la implementación de los algoritmos se emplearon las métricas de matriz de correlación, matriz de confusión, curvas de ROC, precisión, recuperación, puntaje F1, Exactitud. Los resultados mostraron que el algoritmo de RF obtuvo un rendimiento mayor con un AUC de 0.998 en la predicción de campañas de salud.

Agradecimientos

Damos un agradecimiento especial a la empresa privada de venta masiva por ser parte de este proyecto de investigación y al Departamento de Investigación de la Universidad XYZ por la guía brindada en realización del mismo, a través del incentivo xxx.

Referencias

[1] K. Gioergeva. (2022, April 14). Una crisis tras otra: Cómo puede responder el mundo. [Online]. Available: https://meetings.imf.org/es/IMF/Home/News/Articles/2022/04/14/sp041422-curtain-raiser-sm2022

[2] M. Santillán & E. Struyf. (2022, September). La campaña del gobierno peruano para salir de la cuarentena por COVID 19. Un estudio de caso instruccional. [Online]. Available: https://repositorio.ulima.edu.pe/handle/20.500.12724/15458

[3] A. Gutiérrez, "Anuario estadístico sectorial 2022," Ministerio de Trabajo y Promoción del Empleo, vol. 01, July 2023. https://cdn.www.gob.pe/uploads/document/file/4930317/Anuario%202022.pdf?v=1691004485

[4] S. Florian, R. Ruiz, N. Valle & M. Elias, "Situación del mercado laboral en Lima Metropolitana," Instituto Nacional de Estadística e Informática, vol. 08, August 2023.

[5] Instituto Nacional de Estadística e Informática (2020, December). Gastos destinados al sector salud 2007-2021. [Online]. Available: https://www.inei.gob.pe/estadisticas/indice-tematico/health-spending/

[6] Instituto Nacional de Estadística e Informática (2021, December). Población que reportó padecer algún problema de salud crónico según ámbito geográfico, 2012-2022. [Online]. Available: https://m.inei.gob.pe/estadisticas/indice-tematico/health/

[7] E. Orlova, "Innovation in Company Labor Productivity Management: Data Science Methods Application," Applied System Innovation, vol. 4, no. 3, pp. 68, September 2021. https://doi.org/10.3390/asi4030068

[8] K. Norman, L. Burrows, L. Chepulis, R. Keenan & R. Lawrenson, "Understanding weight management experiences from patient perspectives: qualitative exploration in general practice," BMC Primary care , no. 45, February 2023. https://doi.org/10.1186/s12875-023-01998-7

[9] C. Dos Santos, J. Teruel, R. Puppin, E. Yoshio & V. Cortez, "Nutrition Literacy Level in Bank Employees: The Case of a Large Brazilian Company," Nutrients, vol. 15, May 2023. https://doi.org/10.3390/nu15102360

[10] S. Uppal, B. Kansekar, S. Mini & D. Tosh, "Health Dote: A blockchain-based model for continuous health monitoring using interplanetary file system," ScienceDirect, vol. 3, April 2023. https://doi.org/10.1016/j.health.2023.100175

[11] M. Farrona, B. Wipfli, S. Thosar, E. Colino, J. García, L. Gallardo, J. Felipe & J. López, "Effectiveness of worksite wellness programs based on physical activity to improve workers' health and productivity: a systematic review," Systematic Reviews, no. 87, May 2023. https://doi.org/10.1186/s13643-023-02258-6

[12] S. Šćepanović, M. Constantinides, D. Quercia & S. Kim, "Quantifying the impact of positive stress on companies from online employee reviews," Scientific Reports, no. 1603, January 2023. https://doi.org/10.1038/s41598-022-26796-6

[13] B. Szaszi, N. Hajdu, P. Szecsi, E. Tipton & B. Aczel, "A machine learning analysis of the relationship of demographics and social

gathering attendance from 41 countries during pandemic," Nature, no. 724, January 2022. https://doi.org/10.1038/s41598-021-04305-5

[14] K. Kurisu, Y. Song y K. Yoshiuchi, "Developing Action Plans Based on Machine Learning Analysis to Prevent Sick Leave in a Manufacturing Plant," Occupational and Environmental Medicine, vol. 45, no. 2, pp. 140-145, February 2023. https://doi.org/10.1097/JOM.0000000000002700

[15] N. Barrera, R. Torres, J. Rodríguez, O. Espinosa, S. Avellaneda & J. Ramírez, "A recommender system for occupational hygiene services using natural language processing," Healthcare Analytics, vol. 3, November 2023. https://doi.org/10.1016/j.health.2023.100148

[16] P. Nascimento, J. Nunes, J. Lirio, T. Russomano & F. Porto, "Effects of exercises performed in the work environment on occupational stress: A systematic review," Journal of Bodywork and Movement Therapies, vol. 35, pp. 182-189, April 2023. https://doi.org/10.1016/j.jbmt.2023.04.061

[17] B. Lavezo, A. Francisco, a. Machado, S. Sambugaro, G. Lapasini, E. Cardoza & R. Souza, "Data mining in occupational safety and health: A systematic mapping and roadmap," Scielo Brasil, vol. 31, September 2021. https://doi.org/10.1590/0103-6513.20210048

[18] D. Meyer, M. Jayawar, S. Muir, D. Ho & O. Sackett, "Increasing Awareness of the Importance of Physical Activity and Healthy Nutrition: Results From a Mixed-Methods Evaluation of a Workplace Program," Journal of Physical Activity & Health, vol. 14, pp. 259-266, April 2019. https://doi.org/10.1123/jpah.2017-0608

[19] T. Hauth, J. Peiró, J. Mesa & A. Soriano, "Self-perceived Transformational Leadership Decreases Employee Sick Leave, but Context Matters," Revista de Psicologia del Trabajo y de las Organizaciones, vol. 39, no. 1, pp. 37-45, March 2023. https://doi.org/10.5093/jwop2023a5

[20] D. Habr, B. Wolf, K. Schuler & D. Chari, "Patients at the Heart of the Scientific Dialogue: An Industry Perspective," Oncology and Therapy, vol. 11, pp. 15-24, January 2023. https://doi.org/10.1007/s40487-023-00220-z

[21] N. Magnavita, "Workplace Health Promotion Embedded in Medical Surveillance: The Italian Way to Total Worker Health Program," International Journal of Environmental Research and Public Health, vol. 20, no. 4, February 2023. https://doi.org/10.3390/ijerph20043659

[22] D. Lepore, K. Dolui, O. Tomashchuk, H. Shim, C. Puri, Y. Li, N. Chen & F. Spigarelli, "Interdisciplinary research unlocking innovative solutions in healthcare," Technovation, vol. 120, February 2023. https://doi.org/10.1016/j.technovation.2022.102511

[23] V. Silva, B. Gorgulho, D. Marchioni, S. Alvim, L. Giatti, T. Araujo, A. Alonso, I. Santos, P. Lotufo & I. Benseñor, "Recommender System Based on Collaborative Filtering for Personalized Dietary Advice: A Cross-Sectional Analysis of the ELSA-Brasil Study," International Journal of Environmental Research and Public Health, vol. 19, no. 22, November 2022. https://doi.org/10.3390/ijerph192214934

[24] S. Reeja, S. Lavanya, L. Monish & L. Abdul, "It Employee Stress Prediction by Using Machine Learning and Computer Vision Technique," International Journal of Advanced Research in Computer Science, vol. 12, August 2021. https://doi.org/10.26483/ijarcs.v12i0.6728

[25] H. Jade & G. Bulaj, "Health education via "empowerment" digital marketing of consumer products and services: Promoting therapeutic benefits of self-care for depression and chronic pain," Frontiers in Public Health, vol. 10, January 2023. https://doi.org/10.3389/fpubh.2022.949518

[26] A. Pawlicka, M. Pawlicki, R. Tomaszewska, M. Choraś & R. Gerlach, "Innovative machine learning approach and evaluation campaign for predicting the subjective feeling of work-life balance among employees," Plos One, vol. 15, no. 5, May 2020. https://doi.org/10.1371/journal.pone.0232771

[27] N. Chanin. (2020, August 26). El proceso de ciencia de datos. [Online]. Available: https://resources.experfy.com/bigdata-cloud/the-data-science-process/

[28] C. Aracena, F. Villena, F. Arias & J. Dunstan. "Aplicaciones de aprendizaje automatizado en salud," Revista medica clinica los Condes, vol. 33, no. 6, pp. 568-575, October 2022. https://doi.org/10.1016/j.rmclc.2022.10.001

[29] R. Thanki & K. Gmbh. "A deep neural network and machine learning approach for retinal fundus image classification," Healthcare Analytics, vol. 3, November 2023. https://doi.org/10.1016/j.health.2023.100140

[30] K. Tompra, G. Papageorgiou, C. Tjortjis. "Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification," Alforithms, vol.15, no.178, May 2024. https://doi.org/10.3390/a17050178

[31] M. Kumar, D. Dembla, S. Bhatia. "Prediction of Cardiovascular Disease using Machine Learning Algorithms," International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, 2024.https://doi.org/10.14569/IJACSA.2024.0150319