UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TESIS

MODELAMIENTO DE LA SATISFACCIÓN LABORAL DE DOCENTES DE EDUCACIÓN BÁSICA MEDIANTE TÉCNICAS MACHINE LEARNING

PRESENTADA POR:

LUIS ALBERTO HOLGADO APAZA

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PUNO, PERÚ

2022

Reporte de similitud

NOMBRE DEL TRABAJO

MODELAMIENTO DE LA SATISFACCIÓN LABORAL DE DOCENTES DE EDUCACIÓN BÁSICA MEDIANTE TÉCNICAS MACHIN E LE

AUTOR

Luis Alberto Holgado Apaza

RECUENTO DE PALABRAS

30088 Words

RECUENTO DE CARACTERES

152244 Characters

RECUENTO DE PÁGINAS

120 Pages

TAMAÑO DEL ARCHIVO

3.3MB

FECHA DE ENTREGA

Jul 4, 2023 1:33 PM GMT-5

FECHA DEL INFORME

Jul 4, 2023 1:35 PM GMT-5

18% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base

- 15% Base de datos de Internet
- 3% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Cross
- 10% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Coincidencia baja (menos de 12 palabras)

UNA PUNO

Firmado digitalmente por TITO LIPA Jose Panfilo FAU 20145496170 hard
Motivo: Soy el autor del documento
Fecha: 05.07.2023 19:52:04 -05:00

UNA PUNO

Firmado digitalmente por CARPIO VARGAS EDGAR ELOY
Motivo: Soy el autor del documento
Fecha: 04.07.2023 13:38:14 -05:00

Resumen

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TESIS

MODELAMIENTO DE LA SATISFACCIÓN LABORAL DE DOCENTES DE EDUCACIÓN BÁSICA MEDIANTE TÉCNICAS MACHINE LEARNING

PRESENTADA POR:

LUIS ALBERTO HOLGADO APAZA

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

Universidad Nacional del Altiplano
Firmado digitalmente por COYLA IDME Elmer FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 29.06.2023 18:23:00 -05:00

D.Sc. ELMER COYLA IDME

PRIMER MIEMBRO

UNA PUNO
Firmado digitalmente por APAZA TARQUI Alejandro FAU 20145496170 hard
Motivo: Soy el autor del documento
Fecha: 26.06.2023 13:52:50 -05:00

D.Sc. ALEJANDRO APAZA TARQUI

SEGUNDO MIEMBRO

UNA PUNO
Firmado digitalmente por JIMENEZ CHURA ADOLFO CARLOS
Motivo: Soy el autor del documento
Fecha: 21.06.2023 09:29:44 -05:00

D.Sc. ADOLFO CARLOS JIMENEZ CHURA

ASESOR DE TESIS

UNA PUNO
Firmado digitalmente por CARPIO VARGAS EDGAR ELOY
Motivo: Soy el autor del documento
Fecha: 04.07.2023 08:37:10 -05:00

Dr. EDGAR ELOY CARPIO VARGAS

Puno, 28 de diciembre de 2022

ÁREA: Ingeniería de software
TEMA: Modelamiento de la satisfacción laboral de docentes de educación básica mediante técnicas machine learning
LÍNEA: Desarrollo de aplicaciones

DEDICATORIA

A Dios, por darme vida y sabiduría para enfrentar este desafío académico.

A la memoria de mi padre Ricardo.

A mi querida madre Dolores.

A mi amada hija, Luciana Alandra, y a mi hijo amado Luis Alejandro.

A mi compañera de vida Solinka.

Con gratitud y amor,

Luis Alberto Holgado Apaza

AGRADECIMIENTOS

- A los docentes y personal administrativo de la escuela de Posgrado Doctorado-Doctorado en Ciencias de la computación de la Universidad Nacional del Altiplano de Puno. Su invaluable contribución ha sido fundamental para la mejora continua de los profesionales en este campo.

- A mi asesor de tesis por su apoyo incondicional y el tiempo que ha dedicado a esta investigación. Su orientación experta, su sabiduría y su paciencia han sido de un valor incalculable. Sin su guía, este logro no habría sido posible.

- Mi gratitud a los miembros del jurado por su valioso tiempo y por sus aportes significativos a esta investigación. Sus comentarios y sugerencias han enriquecido enormemente este trabajo de investigación.

ÍNDICE GENERAL

Pág.

DEDICATORIA i

AGRADECIMIENTOS ii

ÍNDICE GENERAL iii

ÍNDICE DE TABLAS v

ÍNDICE DE FIGURAS vi

ÍNDICE DE ANEXOS viii

RESUMEN ix

ABSTRACT x

INTRODUCCIÓN 1

CAPÍTULO I

REVISIÓN DE LA LITERATURA

1.1. Marco teórico 3

1.2. Antecedentes 18

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema 24

2.2. Enunciado del problema 25

2.3. Justificación 26

2.4. Objetivos 26

2.4.1. Objetivo general 26

2.4.2. Objetivos específicos 27

2.5. Hipótesis 27

2.5.1. Hipótesis general 27

2.5.2. Hipótesis específicas 27

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. Lugar de estudio 29

3.2. Población 29

3.3. Muestra 29

3.4. Método de investigación 30

3.5. Descripción detallada de métodos por objetivos específicos 30

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Identificar los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica 40

4.2. Establecer el modelo logístico para la satisfacción laboral de docentes de educación básica del Perú 46

4.3. Predecir utilizando árboles de decisión la satisfacción laboral de docentes de educación básica del Perú 55

4.4. Predecir utilizando clasificadores combinados la satisfacción laboral de docentes de educación básica del Perú 62

4.5. Comparar las métricas de la técnica de regresión logística respeto a la técnica de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú. 69

4.6. Discusión de los resultados 74

CONCLUSIONES 76

RECOMENDACIONES 78

BIBLIOGRAFÍA 79

ANEXOS 89

ÍNDICE DE TABLAS

		Pág.
1.	Valoración del área bajo la curva ROC (AUC)	17
2.	Resumen global del conjunto de datos ENDO-2018	33
3.	Descripción de las características seleccionadas	45
4.	Métricas del modelo de regresión logística	47
5.	Valores AUC del modelo de Regresión logística	52
6.	Métricas del modelo de árbol de decisión-CART	57
7.	Valores AUC del modelo de Árbol de decisión	59
8.	Métricas para los modelos de clasificadores combinados	64
9.	Valores AUC del modelo de Bosques aleatorios	66
10.	Resumen estadístico de métricas en el conjunto de datos de entrenamiento	70

ÍNDICE DE FIGURAS

		Pág.
1.	Esquema del proceso KDD.	4
2.	Resumen de algoritmos de aprendizaje automático	7
3.	Esquema general de aprendizaje por refuerzo	15
4.	Estructura de una matriz de confusión para dos clases	16
5.	Metodología propuesta para el modelamiento de la satisfacción laboral de docentes	32
6.	Distribución de clases en la variable objetivo.	38
7.	Script para la obtención de la exactitud en variables categóricas	40
8.	Script para la obtención de la exactitud en variables numéricas	41
9.	Evolución de la exactitud y número de características	41
10.	Script para la obtención de puntuaciones Chi-Cuadrado	42
11.	Puntuaciones Chi-Cuadrado en la predicción de la satisfacción laboral docente	43
12.	Script para la obtención de puntuaciones ANOVA	43
13.	Puntuaciones ANOVA en la predicción de la satisfacción laboral docente	44
14.	Conjunto de datos final	45
15.	Script para la construcción del modelo de regresión logística	46
16.	Matriz de confusión del modelo de regresión logística	46
17.	Área bajo la curva ROC (AUC) del modelo de regresión logística	48
18.	Primera prueba de significancia estadística de variables predictoras	49
19.	Segunda prueba de significancia estadística de variables predictoras	50
20.	Modelo de regresión logística para la predicción de la satisfacción laboral docente	50
21.	Exactitud del modelo de regresión logística	51
22.	Conjunto de datos de mediciones de métricas por modelo	52
23.	Resultados de prueba de normalidad del AUC del modelo de Regresión logística	53
24.	Resultados de la prueba de hipótesis específica 1	54
25.	Script para la construcción del modelo de árbol de decisión	55
26.	Matriz de confusión del modelo de árbol de decisión-CART	56
27.	Área bajo la curva ROC (AUC) del modelo de árbol de decisión	58
28.	Resultados de prueba de normalidad del AUC del modelo de Árbol de decisión	60
29.	Resultados de la prueba de hipótesis específica 2	61
30.	Script para la construcción de los modelos de clasificadores combinados	62
31.	Matriz de confusión de los modelos de clasificadores combinados	63

32. Área bajo la curva ROC (AUC) de los modelos de clasificadores combinados 65
33. Resultados de prueba de normalidad del AUC del modelo de Bosques aleatorios 67
34. Resultados de la prueba de hipótesis específica 3 68
35. Métricas en el conjunto de datos de entrenamiento 70
36. Matriz de confusión de modelos estudiados en el conjunto de datos de prueba 71
37. Métricas obtenidas en el conjunto de datos de prueba 72
38. Comparativa de área bajo la curva ROC (AUC) 73

ÍNDICE DE ANEXOS

Pág.

1. Cuadro de matriz de consistencia 90

2. Conjunto de datos de mediciones de métricas por modelo 93

3. Código fuente para la generación del conjunto de datos de mediciones de métricas por modelo 106

4. Funciones auxiliares para el análisis exploratorio de datos 107

RESUMEN

La satisfacción laboral del docente, es un aspecto importante en el desempeño académico, retención de los estudiantes y retención de los maestros. En el presente estudio se determinó el modelo predictivo de la satisfacción laboral de docentes de educación básica mediante técnicas de aprendizaje automático. El conjunto de datos original estuvo conformado por 15087 instancias y 942 atributos procedentes de la encuesta nacional a docentes de instituciones educativas públicas y privadas de educación básica regular (ENDO-2018) desarrollado por Ministerio de Educación del Perú. Las técnicas de selección de características empleadas fueron el filtro ANOVA F-test y el filtro Chi-Cuadrado. En la fase modelado se emplearon los algoritmos de Regresión logística, Gradient Boosting, Random Forest, XGBoost, Decision Trees-CART. El algoritmo de Random Forest obtiene una exactitud del 73 %, sensibilidad del 74.8 %, AUC del 0.82, menor valor de falsos negativos 163 y mayor valor de verdaderos positivos 484 en la matriz de confusión. Los ingresos económicos, la satisfacción con la vida, con la autoestima, con la actividad pedagógica, con la relación con el director (a), percepción de las condiciones de vida, satisfacción con sus relaciones familiares, problema de salud relacionado con la depresión y la satisfacción de la relación con sus colegas resultaron ser los predictores más importantes.

Palabras clave: Aprendizaje automático, Aprendizaje supervisado, Ciencia de datos, inteligencia artificial, Satisfacción laboral, Modelado predictivo.

ABSTRACT

Teacher job satisfaction is an important aspect of academic performance, student retention, and teacher retention. In the present study, the predictive model of the job satisfaction of basic education teachers was determined using machine learning techniques. The original dataset consisted of 15087 instances and 942 attributes from the national survey of teachers in public and private educational institutions of regular basic education (ENDO-2018) developed by the Ministry of Education of Peru. The feature selection techniques used were the ANOVA F-test filter and the Chi-Square filter. In the modeling phase, logistic regression, Gradient Boosting, Random Forest, XGBoost, Decision Trees-CART algorithms were used. The Random Forest algorithm obtained an accuracy of 73 %, sensitivity of 74.8 %, AUC of 0.82, lower value of false negatives 163 and higher value of true positives 484 in the confusion matrix. Financial income, life satisfaction, self-esteem, teaching activity, relationship with the principal, perception of living conditions, satisfaction with family relationships, health problems related to depression and satisfaction with the relationship with colleagues turned out to be the most important predictors.

Keywords: Machine learning, Supervised learning, Data science, Artificial intelligence, Job satisfaction, Predictive model.

INTRODUCCIÓN

Actualmente nos encontramos en la denominada Cuarta Revolución Industrial o Industria 4.0, donde se cuenta con tecnologías en la nube para el acceso a software, almacenamiento de archivos y procesamiento de datos, permitiéndonos de esta manera la gestión y el análisis de datos masivos con fines de obtener valor, ventaja competitiva, aplicar estrategias empresariales y, principalmente, tomar decisiones guiadas por datos.

El Perú no es ajeno a este fenómeno, consciente de ello se aprobó la "Estrategia Nacional de Datos Abiertos Gubernamentales del Perú 2017-2021" y el "Modelo de Datos Abiertos Gubernamentales del Perú", mediante el Decreto Supremo DS-016-2017-PCM; con ello, se creó el Portal Nacional de Datos Abiertos, donde se encuentran conjuntos de datos (*dataset*) de distintas categorías como: Transporte, agua y saneamiento, salud, gobernabilidad, educación, prosperidad económica, medio ambiente, sociedad de la información entre otros.

Debido a que la satisfacción laboral del docente es uno de los principales factores que afectan el rendimiento, la retención de estudiantes y maestros, es importante abordar esta temática desde diversos enfoques. Por ello, en el presente estudio se emplean técnicas de aprendizaje automático para entender de mejor manera cuales son los factores que afectan la satisfacción laboral del docente. Los factores encontrados permitirán dotar de información a los funcionarios del sector educación, ayudándolo en una toma de decisiones acertadas en políticas educativas, con el objetivo de garantizar una educación de calidad que conlleve al desarrollo del país. Además, en esta investigación se pondrá al alcance de la comunidad científica los resultados de la aplicación de cinco algoritmos de aprendizaje automático en la satisfacción laboral de docentes de educación básica regular.

Motivado en ello, en el presente estudio se propone determinar el modelo predictivo de la satisfacción laboral de docentes de educación básica mediante técnicas de aprendizaje automático. Para ello, se obtienen los datos de la Encuesta Nacional a Docentes (ENDO-2018), desarrollada por el Ministerio de Educación del Perú para identificar los predictores más confiables para la clasificación de la satisfacción e insatisfacción laboral de los docentes desde un enfoque de aprendizaje automático, con el fin de construir un conjunto de datos apropiado para el problema en cuestión. Luego, se realiza la comparación de los algoritmos de regresión logística (LR), Gradient Boosting (GB), Random Forest (RF), XGBoost (XGB) y Decision Trees-CART (DT) para determinar el modelo más adecuado para el caso de estudio.

La presente tesis está organizada de la siguiente forma: En el capítulo I, se explica una serie de conceptos que son necesarios para el entendimiento de la investigación, como: El proceso de descubrimiento de conocimientos en bases de datos (KDD), algoritmos de aprendizaje automático, métricas para la evaluación de modelos predictivos; así también en este capítulo se presenta una revisión de trabajos relacionados al estudio, mismo que se tomaron como referencias. En el capítulo II se muestra la identificación del problema, la justificación, los objetivos a alcanzar y las hipótesis que se pretenden comprobar. En el capítulo III se presenta los materiales y métodos que explicitan los recursos empleados y el procedimiento para el logro de cada uno de los objetivos. En el capítulo IV se muestran los principales resultados y la discusión. Finalmente se presenta las conclusiones y recomendaciones para futuras investigaciones.

CAPÍTULO I
REVISIÓN DE LITERATURA

1.1. Marco teórico

1.1.1. El Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD)

Fayyad et al. (1996) definen a KDD como un proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles, y entendibles en los datos. En este escenario, los datos se refieren a un conjunto de hechos, y los patrones son expresiones en algún lenguaje que describen los datos de manera compacta. Con respecto al término proceso esto implica que el KDD comprende muchos pasos o fases, entre los que se encuentran: la preparación de datos, búsqueda de patrones, evaluación del conocimiento encontrado, y refinamiento; los mismos que pueden ser repetidos en múltiples iteraciones. Por no trivial, debe entenderse que alguna búsqueda o inferencia es llevada a cabo; es decir, involucra la búsqueda de estructuras, modelos, patrones o parámetros. Los patrones descubiertos deben ser válidos sobre nuevos datos con algún grado de certeza, para que puedan describir y/o predecir confiablemente el comportamiento futuro de alguna entidad.

El proceso de KDD es interactivo e iterativo, involucrando numerosos pasos con muchas decisiones tomadas por el usuario. La siguiente figura presenta el proceso de KDD:

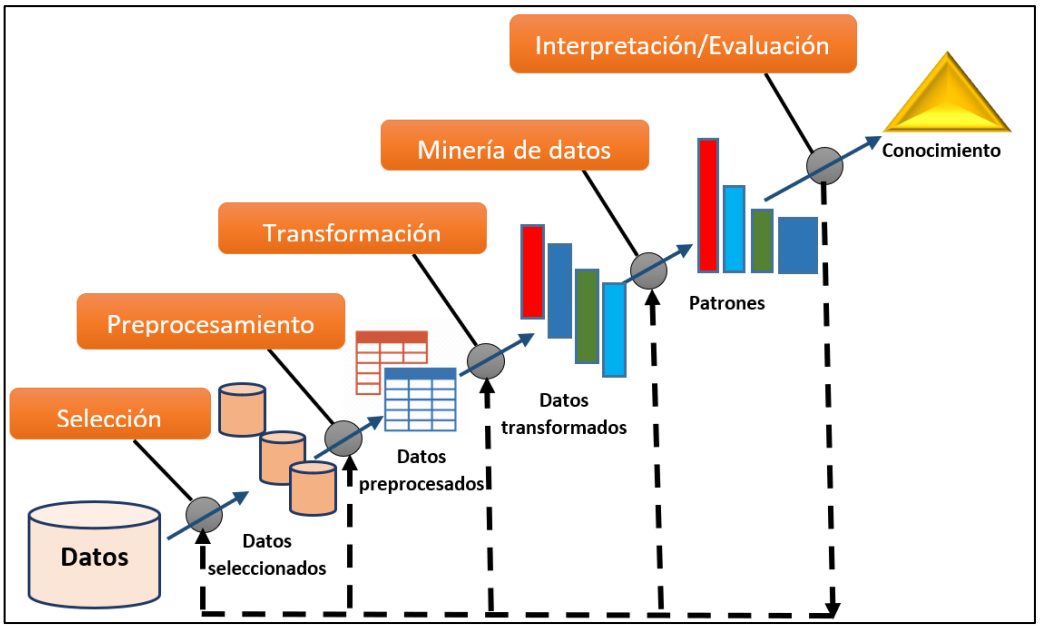


Figura 1. Esquema del proceso KDD.

Fuente: Adaptado de (Fayyad *et al.*, 1996)

A continuación se detalla cada una de las fases realizadas en base a Pérez y Santín (2008):

En la fase de selección, donde se integran y recopilan los datos, determinamos las fuentes de información útiles y dónde conseguirlas, identificamos y seleccionamos las variables más relevantes en los datos, y también se aplican técnicas de muestreo adecuadas.

En la fase de preprocesamiento, se realiza la exploración mediante técnicas de análisis exploratorio de los datos, buscando su distribución, simetría, normalidad y las correlaciones existentes entre los mismos; otra actividad dentro de esta fase es la limpieza de los datos, dado que pueden existir datos atípicos, valores faltantes y valores erróneos.

La fase de transformación de los datos, se lleva a cabo mediante técnicas de reducción o aumento de la dimensionalidad, escalado simple y multidimensional.

En la fase de minería de datos o modelado, determinamos la tarea a realizar (clasificación, regresión, agrupamiento) y de acuerdo a ello elegimos la técnica a emplear.

Finalmente, en la fase evaluación/interpretación, se avalúan los patrones y estos son analizados por expertos del negocio. En esta fase, de ser necesario se vuelve a las fases anteriores para una nueva iteración.

1.1.2. Minería de datos

La minería de datos emplea técnicas de reconocimiento de patrones para luego extraer estos patrones y tendencias relevantes de los datos (Van Loo et al., 2020), las técnicas más empleadas son las descriptivas y las predictivas.

Llegado a este punto, es posible definir a la minería de datos como el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, datos almacenes, la Web, otros repositorios de información o datos que se transmiten a sistema de forma dinámica (Han *et al*., 2012).

1.1.3. Inteligencia artificial (IA)

Se define como una rama de las ciencias de la computación que se ocupa de la compresión, desde el punto de vista informático, de lo que comúnmente se denomina comportamiento inteligente.

Incluye distintas ramas como el aprendizaje automático (*machine learning*), el procesamiento de lenguaje natural, los sistemas expertos, la visión artificial, etc., y es la base de otros muchos como la robótica o el *big data* dos de las áreas que más están creciendo actualmente (Rainer y Rodríguez, 2017, p. 17). Por su parte Pazos *et al.* (2007) afirman que la IA se encarga del estudio de la inteligencia en elementos artificiales y desde el punto de vista de la ingeniería, propone la creación de elementos que posean un comportamiento inteligente (p. 10).

1.1.4. Aprendizaje automático (Machine learning)

Sub campo de la Inteligencia artificial conocido también como aprendizaje de máquina, Ponce *et al.* (2014) afirman que el aprendizaje automático corresponde a programas computacionales que buscan optimizar los parámetros de un modelo usando datos previos o datos de entrenamiento. Los modelos pueden ser inductivos, cuando permiten hacer predicciones sobre el futuro o bien descriptivos cuando permiten generar conocimiento a partir de los datos (p. 88).

Por su parte Russo *et al.* (2016) afirman que el aprendizaje automático es un área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible este aprendizaje mediante el entrenamiento con grandes volúmenes de datos. En la actualidad existen diferentes modelos que utilizan esta técnica y consiguen una precisión incluso superior a la de expertos en las mismas tareas, por ejemplo, en el reconocimiento de objetos en una imagen, detección de fraude, diagnósticos médicos, reconocimiento de voz, entre otras aplicaciones. El aprendizaje automático y la inteligencia artificial son técnicas innovadoras disruptivas que prometen cambiar la sociedad tal y como la conocemos, estos cambios tendrán su reflejo en el mundo laboral en general (Beunza-Nuin *et al.*, 2020).

Los algoritmos de aprendizaje automático se clasifican en aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

Aprendizaje supervisado

De acuerdo con Moreno *et al.* (2001) estos algoritmos predicen el valor de un atributo denominado etiqueta o target de un conjunto de datos, conocidos otras variables o atributos (descriptivos o predictores). A partir de un conjunto de datos cuya etiqueta se conoce se procede a inducir a una relación entre dicha etiqueta y otra serie de atributos. Estas relaciones servirán para poder realizar la predicción en datos cuya etiqueta es aún no conocida. A esta forma de trabajo se conoce como aprendizaje supervisado y se desarrolla en dos fases: Fase de entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y la fase de prueba (prueba del modelo sobre el resto de los datos). Se puede incluir en estas técnicas todos los tipos de regresión, series de tiempo, análisis de la varianza, y la covarianza, análisis discriminante, redes neuronales, algoritmos genéticos, árboles de decisión y técnicas bayesianas.

Por su parte Rosado *et al.* (2015) afirman que las técnicas predictivas tienen las tareas de clasificación y regresión. Las tareas de regresión persiguen la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión) (Valcárcel, 2004). En relación con la clasificación Aluja (2001) menciona que si la respuesta es categórica (p. e. la compra o no de un producto) se dice que se trata de un problema de clasificación. A continuación, se muestran los algoritmos de aprendizaje automático más utilizados de acuerdo con (Gironés *et al.*, 2017):

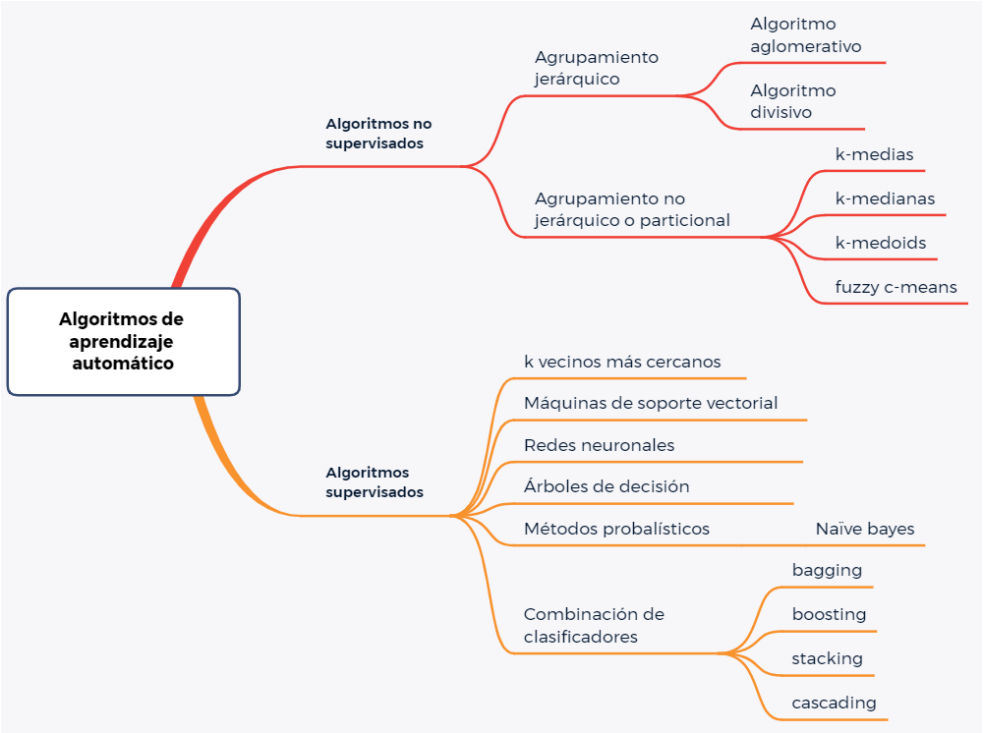


Figura 2. Resumen de algoritmos de aprendizaje automático
Fuente: Adaptado de (Gironés et al., 2017)

El modelo de regresión logística

Es un tipo especial de regresión que es utilizado cuando se pretende predecir y explicar una variable categórica binaria, en contraste de una variable dependiente numérica o métrica. La ventaja de esta técnica es la de verse menos afectada que el análisis discriminante cuando no se cumplen los supuesto básicos en concreto la normalidad de las variables, sumado a ello permite utilizar variables no métricas por medio de codificación (Hair *et al.*, 1999).

Pérez (2008) menciona que para una única variable independiente X, el modelo de regresión logística toma la forma de:

$$\ln\left(\frac{p}{q}\right) = \alpha_0 + \alpha_1 X$$

Donde:

α_0 y α_1 son constantes y X es la variable que puede ser aleatoria o no, continua o discreta. Este modelo puede ser fácilmente generalizado para k variables dependientes, dando lugar al modelo de regresión logístico múltiple que se expresa a continuación:

$$\ln\left(\frac{p}{q}\right) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k$$

Es posible escribir el modelo logístico de la siguiente manera:

$$\ln\left(\frac{p}{q}\right) = \alpha_0 + \alpha_1 X \leftrightarrow \ln\left(\frac{p}{p-1}\right) = \alpha_0 + \alpha_1 X \leftrightarrow \frac{p}{p-1} = e^{\alpha_0+\alpha_1 X} \leftrightarrow p = \frac{e^{\alpha_0+\alpha_1 X}}{1+e^{\alpha_0+\alpha_1 X}} \leftrightarrow p = \frac{1}{1+e^{-(\alpha_0+\alpha_1 X)}}$$

A la función: $f(z) = \frac{1}{1+e^{-z}}$ denominamos función logística.

El modelo de regresión logística múltiple tendrá la expresión:

$$p = \frac{1}{1+e^{-(\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\cdots+\alpha_k X_k)}}$$

Arboles de decisión

Los árboles de decisión permiten realizar tareas de clasificación de los datos en grupos basados en los valores de las variables. El mecanismo principal consiste en elegir un atributo como raíz y construir el árbol según las variables más significativas (Pérez y Santín, 2008, p. 8).

Cuando la variable dependiente u objetivo (target) es categórica, se los llama árboles de clasificación; por otro lado, si la variable dependiente es continua, estaríamos hablando de árboles de regresión. Ambos tipos de árboles de decisión dan el nombre común CART (*Classification and Regression Trees*) (Breiman *et al.*, 1987; Gironés *et al.*, 2017).

A continuación, se muestra el algoritmo genérico para construir un árbol de decisión:

Algoritmo 1: Árbol de decisión

Entrada: Conjunto de datos a clasificar D

$T = \{D\}$ // El árbol inicial es un solo nodo hoja

Etiquetar T de acuerdo a C_c

$p = \{T\}$ // Lista de nodos pendientes

mientras no se deba para según C_p **hacer**:

Seleccionar un nodo q de acuerdo a C_s

si es posible particionar q según C_d **entonces**

Particionar q en $q_1, \dots, q_P$

Etiquetar $q_1, \dots, q_P$ segun C_c

Añadir $q_1, \dots q_P$ a p

Sustituir q en T por un nodo interno

fin si

Eliminar q de p

fin mientras

retornar T

fin Algoritmo

Classification and Regression Trees (CART)

Este algoritmo permite la generación árboles de decisión binarios, lo que quiere decir que cada nodo del árbol se divide en exactamente dos ramas. La ventaja de este algoritmo es que admite variables de entrada y de salida nominales, ordinales y continuas, lo que permite resolver problemas de clasificación y de regresión (Parra, 2019). Este algoritmo usa el índice de Gini como una medida de selección de atributos para construir un árbol de decisión:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C/A_{ij})$$

Siendo, $G(C/A_{ij})$ es igual a:

$$G(C/A_{ij}) = -\sum_{j=i}^{M_i} p(C_k/A_{ij})(1 - p(C_k/A_{ij}))$$

Donde:

A_{ij} es el atributo empleado para ramificar el árbol

J es el número de clases

M_i es número de valores distintos que tiene el atributo A_i

$p(A_{ij})$ es la probabilidad de que A_i tome su j-ésimo valor

$p(C_k/A_{ij})$ es la probabilidad de que una instancia sea de la clase C_k cuando su atributo $A_i$toma si j-ésimo valor.

Random Forest (Bosques aleatorios)

Se trata de un conjunto de árboles que han sido creados mediante un procedimiento aleatorio (Gironés *et al*., 2017). Este algoritmo utiliza el conjunto de datos de entrenamiento T, para luego crear k muestras mediante la técnica de *bootstrap* T_k. Con estas muestras se construyen los árboles $h(x, T_k)$ y el promedio de ellos será el predictor *bagget* en el caso de regresión y el más botado para el caso de clasificación. En adelante para cada (y, x) de T se construyen los árboles en cada T_k que no contienen a (y, x), esto son las muestras que quedaron fuera de las muestras *Bootstrap* (Villa *et al*., 2016). A continuación, se muestra el algoritmo propuesto por Breiman (2001):

Algoritmo 2: Random. Forest

para $i = 1$ **hasta** c **hacer**:

Muestree aleatoriamete los datos de entrenamiento D con reemplazo para producir D_i

Cree un nodo raiz, N_i que contenga D_i

Call BuilderTree(N_i)

fin para

BuildTree(N):

si N contiene instancias de solo una clase **entonces**

retornar

si no

Selecciones aleatoriamente x% de los posibles divisiones de características en N

Seleccione la función F con la ganancia de información más alta para dividir

Cree f nodos secundarios de $N, N_1 \dots N_f$, donde F tiene f posibles valores ($F_1 \dots F_f$)

para $i = 1$ **hasta** f **hacer**

Establezca el contenido de N_i en D_i, donde D_i son todas las instancias en N que coinciden con F_i

Call BuildTree(N_i)

fin para

fin si

fin Algoritmo

Gradient Boosting (Potenciación del gradiente)

De acuerdo con Friedman (2001) este es un algoritmo de optimización numérica que tiene como objetivo minimizar la función de pérdida. Dicho algoritmo agrega iterativamente en cada paso un nuevo árbol de decisión también llamado aprendiz débil que reduzca mejor la función de pérdida (Touzani *et al*., 2018).

En este algoritmo, se ajusta un primer aprendiz débil f_1, el cual predice la variable respuesta y, luego se calculan los residuos $y - f_1(x)$. Se prosigue ajustando un nuevo

modelo f_2 que intenta predecir los residuos del modelo anterior, es decir, trata de corregir los errores cometidos por el modelo f_1. Esto es:

$$f_1(x) \approx y$$

$$f_2(x) \approx y - f_1(x)$$

En una siguiente iteración, se calculan los residuos de los dos modelos anteriores de forma conjunta $y - f_1(x) - f_2(x)$. Se ajusta un tercer modelo f_3 para corregir los errores cometidos por f_1 y f_2.

$$f_3(x) \approx y - f_1(x) - f_3(x)$$

Este proceso se repite M veces, de manera que cada nuevo modelo minimice los residuos del anterior. A continuación, se muestra el algoritmo de potenciación del gradiente propuesto por Friedman (2001):

Algoritmo 3: Gradient Boosting

$F_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$

para $m = 1$ **hasta** M **hacer**:

$$\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1(x)}}, i = 1, N$$

$$a_m = argmin_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i; a]^2$$

$$\rho_m = argmin_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m))$$

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m)$$

fin para

fin Algoritmo

XG Boost (Extreme Gradient Boosting)

Es una técnica que forma parte del aprendizaje supervisado que está basado en árboles de decisión y que en la actualidad constituye el estado del arte en la evolución de la familia de algoritmos de *Boosting* (Espinosa-Zúñiga y Espinosa-Zúñiga, 2020). Este algoritmo fue publicado por primera vez por Tianqi-Chen y Carlos Guestrín en el 2016

(Tianqi Chen y Guestrin, 2016), el cual ha sido optimizado y mejorado continuamente (Li *et al*., 2019). A continuación, se muestra el algoritmo *XGBoost* genérico:

Algoritmo 4: XGBoost

Entrada: conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^{N}$, una función de pérdida diferenciable $L\big(y, F(x)\big)$, un número de aprendices débiles M y una tasa de aprendizaje α.

Inicialice el modelo con un valor constante:

$$\hat{f}_{(0)}(x) = \arg_{\theta}\min \sum_{i=1}^{N} L(y_i, \theta)$$

para $m = 1$ **hasta** M **hacer**:

Calcule los gradientes y hessianas

$$\hat{g}_m(x) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$

$$\hat{h}_m(x) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$

Ajuste el aprendiz debil usando el conjunto de entrenamiento $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ resolviendo el siguiente problema de optimización:

$$\widehat{\emptyset}_m = \arg_{\emptyset\in\Phi}\min \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \emptyset(x_i)\right]^2$$

$$\hat{f}_m(x) = \alpha\widehat{\emptyset}_m(x)$$

Actualizar el modelo:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x)$$

fin para

retornar $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_{(m)}(x)$

fin Algoritmo

Aprendizaje no supervisado

Los métodos no supervisados (*unsupervised methods*) son algoritmos que fundamentan su proceso de entrenamiento en un conjunto de datos sin etiquetas o

clases previamente definidas. En estos métodos, a priori no se conoce ningún valor de la variable objetivo o de clase, ya sea categórico o numérico.

El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas *clustering* o segmentación, donde su objetivo es encontrar grupos similares en los juegos de datos (Gironés *et al.*, 2017, p. 39).

De acuerdo con Hair et al. (1999) los algoritmos para la obtención de conglomerados (*clustering*) más utilizados pueden clasificarse en dos categorías generales: (1) jerárquicos y (2) no jerárquicos. (p. 510).

Aprendizaje por refuerzo

En inglés *reinforcement learning*. Este tipo de aprendizaje automático, recoge los fundamentos de la psicología conductista y teorías educativas como el condicionamiento operante. Una de las leyes más claras que guardan relación con este tipo de aprendizaje es en psicología la ley del afecto. Según Panadero y Alonso-Tapia (2014), "si una conducta tiene un efecto positivo se repetirá cuando se den condiciones similares a aquellas en que el efecto se ha conseguido, mientras que si el efecto es negativo ocurrirá lo contrario". En relación al aprendizaje automático Raschka y Mirjalili (2017), afirman que:

> En el aprendizaje por refuerzo, el objetivo es desarrollar un sistema (agente) que mejore su desempeño en función de las interacciones con el entorno. Dado que la información sobre el estado actual del entorno normalmente también incluye una llamada señal de recompensa, podemos pensar en el aprendizaje por refuerzo como un campo relacionado con el aprendizaje supervisado. Sin embargo, en el aprendizaje por refuerzo, esta retroalimentación no es la etiqueta o el valor de verdad fundamental correcto, sino una medida de qué tan bien se midió la acción mediante una función de recompensa. A través de su interacción con el entorno, un agente puede utilizar el aprendizaje por refuerzo para aprender una serie de acciones que maximizan esta recompensa mediante un enfoque exploratorio de prueba y error o una planificación deliberativa. (p. 46)

A continuación, se muestra un esquema general de aprendizaje por refuerzo donde el agente intenta maximizar la recompensa mediante una serie de interacciones con el entorno. Cada estado puede relacionarse con una recompensa positiva o negativa, y

una recompensa se puede definir como el logro de un objetivo general, como ganar o perder una partida de ajedrez.

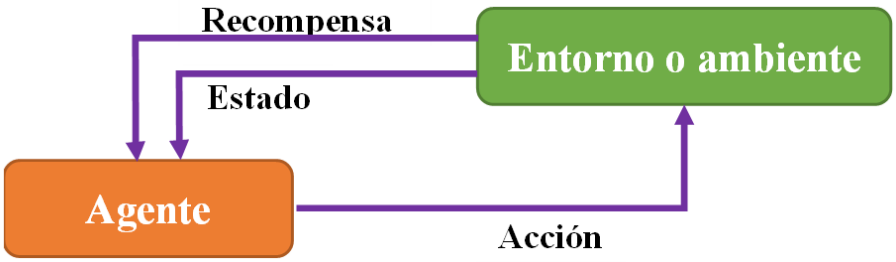


Figura 3. Esquema general de aprendizaje por refuerzo

1.1.5. Metodología para el modelado predictivo

Existen metodologías bien conocidas que proporcionan un enfoque estructurado para desarrollar un modelo basado en datos (DDM) (Fisher *et al.*, 2020). A continuación, mencionamos las más usadas en los ámbitos académicos y laborales: El proceso de descubrimiento de conocimiento en bases de datos (KDD) que permite extraer modelos y patrones de interés en grandes bases de datos (Fayyad y Stolorz, 1997), el CRISP-DM (*Cross Industry Standard Process for Data Mining*), que "es el estándar de facto y un modelo de proceso independiente de la industria para aplicar proyectos de minería de datos" (Schröer et al., 2021) y muestrear, explorar, modificar, modelar y evaluar (SEMMA) (Silva *et al.*, 2020).

1.1.6. Métricas de evaluación para modelos predictivos

Matriz de confusión

Es una tabla de doble entrada, donde se muestra la clasificación observada o real y la clasificación predicha (mediante el clasificador propuesto) para las distintas clases de la variable objetivo. En la Figura 4 se observa la estructura de la dicha matriz para dos clases.

		Clase predicha	
		Negativo	Positivo
Clase real	Negativo	Verdaderos negativos (VN)	Falsos positivos (FP)
	Positivo	Falsos negativos (FN)	Verdaderos positivos (VP)

Figura 4. Estructura de una matriz de confusión para dos clases

Procedemos a detallar algunas métricas comunes empleados en la evaluación de modelos de clasificación a partir de la matriz de confusión.

Exactitud. Representa la proporción de predicciones correctas. La fórmula para obtener este valor la siguiente:

$$Exactitud = \frac{(VN + VP)}{(VN + VP + FN + FP)}$$

Sensibilidad. Representa la proporción de positivos reales que el modelo predijo correctamente, para el presente estudio se considera (Clase positiva 1: Insatisfecho). La fórmula siguiente permite calcular este valor:

$$Sensibilidad = \frac{VP}{(VP + FN)}$$

Especificidad. Representa la proporción de negativos reales que el modelo predijo correctamente, para el presente estudio se considera (Clase negativa 0: Satisfecho).

$$Especificidad = \frac{VN}{(VN + FP)}$$

Valor predictivo positivo (VPP). También llamado precisión, representa la probabilidad de una instancia sea de la clase positiva habiendo sido predicho como positivo por el modelo.

$$VPP = Precision \frac{VP}{(VP + FP)}$$

Valor predictivo negativo (VPN). Representa la probabilidad de una instancia sea de la clase negativa habiendo sido predicho como negativo por el modelo.

$$VPN = \frac{VN}{(VN + FN)}$$

Puntuación F1 (F1-Score). Se define como una media armónica de y la precisión y sensibilidad, donde una puntuación F1 alcanza su mejor valor en 1 y su peor puntuación en 0.

$$F1 = 2 * \frac{Precision * Sensibilidad}{(Precision + Sensibildiad)}$$

Área bajo la curva ROC (AUC). Nos indica que tan bueno es el modelo para discriminar instancias de la clase positiva y la clase negativa. Este valor fluctúa entre 0 y 1. De acuerdo con Cerda y Cifuentes (2012), a medida que este valor se acerca a 1 mayor será la capacidad discriminativa del modelo, entendiéndose como capacidad discriminativa a la habilidad del modelo para distinguir instancias de la clase positiva versus instancias de la clase negativa.

De acuerdo con Gironés *et al.* (2017), se presenta la tabla para la valoración de esta métrica.

Tabla 1

Valoración del área bajo la curva ROC (AUC)

Valor AUC	Interpretación
[0.50-0.60)	Test malo
[0.60-0.75)	Test regular
[0.75-0.90)	Test bueno
[0.90-0.97)	Test muy bueno
[0.95-1.00)	Test excelente

1.1.7. Satisfacción laboral docente

Según Robbins (2004), la satisfacción laboral o satisfacción con el trabajo, se refiere a la actitud general que tiene el individuo hacia su trabajo; también, el autor menciona que una persona con una gran satisfacción con el trabajo tiene actitudes positivas, mientras que las personas insatisfechas poseen actitudes negativas. (p. 72)

En relación a la satisfacción laboral del docente, se puede definir como las reacciones afectivas de los docentes hacia su trabajo o a su función de docente (Skaalvik y Skaalvik, 2011).

1.2. Antecedentes

A continuación, se procede a mencionar los trabajos previos de carácter internacional que se relacionan al presente estudio:

El estudio presentado por Rustam *et al.* (2021), donde se buscó evaluar cinco algoritmos de aprendizaje automático como son: *Random Forest*, Regresión Logística, Máquina de Soporte Vectorial, *Gradient boosting* (GB) y *extreme GB* (EGB), en la predicción de la satisfacción laboral, los investigadores emplearon el conjunto de datos que contenía las reseñas de texto de los empleados de Google, Facebook, Amazon, Microsoft y Apple. En el estudio los autores emplearon técnicas de selección de característica como: frecuencia inversa de términos (TF-IDF), bolsa de palabras (BoW) y vectores globales (GloVe). Los resultados indican que la técnica de selección de características TF-IDF permitió a los algoritmos de regresión logística, bosques aleatorios y extreme GB (EGB) obtener una exactitud del 78 %; así también, a los algoritmos Gradient boosting y Máquina de soporte vectorial obtener una exactitud del 77 %. Finalmente, los autores implementaron un perceptrón multicapa con un 83 % de exactitud.

El estudio presentado por Chen *et al.* (2021), que tuvo como propósito establecer un modelo predictivo para la satisfacción laboral de los trabajadores de la construcción, los autores emplearon algoritmos de árboles de decisión CART (*Classification and Regression Tree*) y redes neuronales. El conjunto de datos estuvo conformado por 280 casos de datos empíricos. Los resultados demuestran que el modelo CART obtuvo una exactitud del 76.15 %, frente a un 71.70 % de la red neuronal.

Saleh y Abu-Soud (2021), afirman que medir la satisfacción laboral de los trabajadores constituye un problema en las organizaciones de gran prestigio, esto debido a que la renuncia repentina de un trabajador podría ocasionar pérdidas en la organización. En razón a ello, presentaron dos técnicas para predecir la satisfacción laboral en Jordania, la de Redes neuronales artificiales y el árbol de decisión J48 implementados en el programa Weka. El conjunto de datos para los experimentos se obtuvo de un cuestionario en línea.

El estudio presentado por Moon *et al.* (2021), tuvo como objetivo determinar las características más significativas de alguien que está deprimido o feliz con su trabajo, y la proporción de hombres y mujeres deprimidos en sus respectivos sectores laborales. Los autores emplearon algoritmos como análisis factorial, bosques aleatorios de clasificación, bosques aleatorios de regresión, *Naive Bayes* y el clasificador K-Vecinos. Los datos empleados fueron recopilados de varias agencias gubernamentales de Bangladesh. Los hallazgos indican que los aspectos más relacionados con las variables depresión y la satisfacción laboral son la edad, la ocupación, el tipo de casa y el deseo de cambio de trabajo; además, los algoritmos de bosques aleatorios y *Naive Bayes* obtuvieron una precisión del 99 %, mientras que el clasificador K Vecinos un 97 %.

En el estudio presentado por Hossen *et al.* (2021), se afirma que la partida de un empleado experto puede crear un problema para una empresa. Este tipo de sucesos están aumentando a nivel mundial, debido a factores como la gran carga laboral, bajos salarios, baja satisfacción laboral y el mal ambiente de trabajo. Los autores realizaron la predicción de la rotación de empleados con ayuda del clasificador de aprendizaje automático. En el estudio, se empleó el algoritmo de selección secuencial (SBS) para la reducción de características, Chi-cuadrado y bosques aleatorios para determinar las variables más importantes en la predicción. Los resultados demuestran que el algoritmo de bosques aleatorios con validación cruzada de 10 veces alcanzó una precisión del 99.4 %.

El estudio presentado por Arambepola y Munasinghe (2021), tuvo como objetivos: estudiar los enfoques disponibles para predecir la satisfacción laboral, identificar los principales factores que influyen en la satisfacción laboral de los profesionales de TI y explorar las posibilidades de generalizar los modelos de predicción de la satisfacción laboral a otros ámbitos de la industria de TI. Los algoritmos empleados fueron *Random Forest*, Regresión Logística, Máquinas de soporte vectorial y Redes neuronales. El conjunto de datos fue obtenido de la encuesta para desarrolladores de Stack Overflow y el conjunto de datos de análisis de recursos humanos de IBM. Los resultados demuestran que el algoritmo de *Random Forest* obtiene los mejores valores para la exactitud del 0.80, precisión de 0.74, especificidad del 0.80 y una puntuación F1 de 0.75.

En el estudio presentado por Pratt *et al.* (2021), se comparó la capacidad predictiva de los algoritmos de regresión logística, bosques aleatorios, K-Vecinos más cercanos y máquinas de soporte vectorial en la satisfacción laboral. El conjunto de datos que emplearon para los experimentos estuvo conformado por los datos de 102 personas. Tras la comparación de los cuatros algoritmos, los resultados evidencian que los algoritmos de bosques aleatorios y regresión logística obtuvieron valores de exactitud más altos de 95.24 % y 80.95 % respectivamente; además, las variables más relevantes para la satisfacción de los empleados, extraídas mediante el algoritmo de bosques aleatorios, están asociadas con el reconocimiento (REC), los buenos sentimientos acerca de la organización (GFO), la alta dirección eficaz (SMG) y la supervisión eficaz (SPV) que constituyen factores intrínsecos de la satisfacción laboral.

En el estudio presentado por Yoo y Rho (2020), el propósito fue identificar los predictores más importantes de la satisfacción laboral de los profesores coreanos a través del aprendizaje automático. Para lograr este objetivo emplearon la técnica de regresión penalizada denominada *Group Mnet*. El estudio permitió identificar 18 predictores de 558, que incluían variables relacionadas con el clima escolar colaborativo, autoeficacia docente, retroalimentación de los maestros y barreras percibidas para el desarrollo profesional.

En el estudio presentado por Talingting (2019), se afirma que uno de los problemas existentes en todas las organizaciones, en especial en las escuelas son la proliferación de los datos y ¿cómo estos datos serán de ayuda para los programas de intervención y toma de decisiones? En razón a esto se realizó un estudio que tuvo como propósito mostrar la efectividad de análisis de la minería de datos en la predicción de la satisfacción laboral de los administradores de las escuelas. El conjunto de datos para los experimentos fue obtenido mediante una encuesta elaborada por el departamento de educación de Surigao del Norte de Caraga en Filipinas. Los algoritmos empleados fueron: *Naive Bayes*, árbol de decisión C4.5 y K-vecinos más cercanos. Los resultados con relación a la exactitud son del 80,89 %, 74,52 % y 71,97 % utilizando los algoritmos C4.5, Naive Bayes y K-vecinos más cercanos, respectivamente.

Khera y Divya (2019) desarrollaron un modelo para predecir la deserción de los empleados y brindar a las organizaciones oportunidades para la mejora de la retención. El modelo se desarrolló empleando la técnica de aprendizaje automático denominado

máquina de soporte vectorial (SVM). Los datos empleados para los experimentos se obtuvieron de las oficinas de recursos humanos de tres empresas de TI en la India, que incluían 22 variables de entrada y la situación laboral (variable objetivo). Los resultados de la precisión de la matriz de confusión alcanzaron un 85 %.

En el estudio presentado por Tomás *et al.* (2019), el objetivo fue poner a prueba un modelo integrador y explicativo de la satisfacción laboral de una muestra representativa de docentes en la República Dominicana, mediante un modelo de ecuaciones estructurales. Los resultados demuestran que la capacidad predictiva del *burnout*, el *engagement* y el clima laboral son de un 82.9 % sobre la satisfacción laboral. Los autores concluyen que el contexto escolar en el que se desarrolla la labor docente no solo tiene un efecto en la satisfacción laboral de los docentes y en su calidad de vida, sino también en la calidad educativa del alumnado.

Sisodia *et al.* (2018), presentan un estudio con el objetivo de construir un modelo que permita predecir la tasa de abandono de los empleados a partir del conjunto de datos de análisis de recursos humanos obtenidos del sitio web de Kaggle. Los autores emplearon cinco algoritmos de aprendizaje automático, como la máquina de soporte lineal, árbol de clasificación C5.0, bosques aleatorios, k-vecinos más cercanos y el clasificador *Naive Bayes*. Los resultados demuestran que los algoritmos que presentaron valores más altos de precisión fueron: el de bosque aleatorio con un valor de 0.98, el algoritmo de árbol de clasificación C5.0 con un valor de 0.97 y el de k-vecinos más cercanos con un 0.96.

Jain y Nayyar (2018), mencionan que la rotación de los empleados ocasiona efectos adversos a la organización, por ello las organizaciones están empleando técnicas de aprendizaje automático para predecir la rotación de empleados. En este estudio los autores se propusieron predecir la deserción de los empleados utilizando un enfoque de aprendizaje automático. El conjunto de datos empleado para los experimentos del estudio es el denominado *IBM HR Dataset: Exploratory Data Analysis* que se encuentra en línea. El algoritmo empleado fue el denominado XGBoost. Los resultados indican que la precisión alcanzada por este algoritmo fue muy cercana a 90%; además, en el análisis exploratorio de datos los autores evidencian que los niveles más altos de deserción se observaron en menores niveles de satisfacción laboral.

El estudio presentado por Kuzey (2018), tuvo como objetivo encontrar factores clave que contribuyan a la satisfacción laboral entre los trabajadores de la salud. Para lograr este objetivo, se emplearon las técnicas de Análisis Factorial Exploratorio y Máquinas de Soporte Vectorial (SVM). De acuerdo con el modelo propuesto, los principales factores que contribuyen con la satisfacción laboral son: actitud de los directivos, pago/recompensa, seguridad en el trabajo y los compañeros de trabajo.

En el trabajo presentado por Robertson y Kee (2017), se realiza el estudio de la satisfacción laboral en el lugar del trabajo en el contexto mediado por la computadora. El propósito del estudio fue determinar la asociación entre la satisfacción laboral de un empleado con la cantidad de tiempo que pasa en Facebook interactuando con sus compañeros de trabajo. El conjunto de datos estuvo conformado por los datos de 512 trabajadores de organizaciones y empresas locales del sur de California, EE.UU. La técnica estadística empleada fue la de correlación rho de Spearman. Los resultados indican que la satisfacción laboral de los empleados se asocia positivamente con la cantidad de tiempo que pasan en Facebook interactuando con sus compañeros de trabajo, con un *rho* = 0.134 y p = 0.009.

Otro estudio presentado por Hong-Hua *et al.* (2016) manifiesta que para los docentes, los bajos niveles de satisfacción laboral están relacionado con una mayor intención de dejar o cambiar de carrera. Por ello, realizaron un estudio con el propósito de encontrar los factores que influyen en la satisfacción laboral de docentes de las escuelas primarias de China. Los autores emplearon la técnica denominada Modelos Jerárquicos Lineales o Modelos Multinivel. Los resultados demuestran que los factores que afectan la satisfacción laboral de los docentes del más importante al menos importante fueron: el apoyo al desarrollo profesional, liderazgo instruccional del director, preferencias ocupacionales y compromiso laboral.

El estudio presentado por Park *et al.* (2016) tuvo como propósito investigar el poder predictivo de la experiencia personal, satisfacción con la vida, preparación vocacional y entorno laboral en la satisfacción laboral de adultos con discapacidad en Corea del Sur. Los autores emplearon la técnica de ecuaciones estructurales. El conjunto de datos estuvo conformado por una muestra de 417 personas con discapacidad que están empleados y trabajan en lugares de trabajo asalariados. Los resultados demuestran que la satisfacción con la vida y el entorno laboral, son factores significativos en la

predicción de la satisfacción laboral con β=1.198, p<0.000; β =0.334 y p<0.000, respectivamente.

En el estudio presentado por Malander (2016), se determinó si la satisfacción laboral y algunas variables sociodemográficas y laborales constituyen un antecedente o predictor del síndrome de burnout. Para lograr este objetivo los autores emplearon un modelo de regresión lineal múltiple. El conjunto de datos para la experimentación estuvo conformado por una muestra de 123 docentes de colegios de gestión privada. Los resultados indican que la satisfacción laboral predice significativamente el burnout $F_{(2, 120)}$ =36.46 con un p-valor =0.000. La variancia explicada se encontró en torno al 38%. Los autores concluyen que los factores intrínsecos de la satisfacción laboral fueron los mejores predictores del cansancio emocional, la despersonalización y la realización personal.

El estudio presentado por Nader *et al.* (2014), tuvo como objetivo determinar si la percepción del clima social, la descripción del trabajo (significación, responsabilidad y conocimiento de resultados) y la experiencia de fluidez (Flow) permiten la predicción de la satisfacción y el bienestar en el trabajo. El conjunto de datos para los experimentos estuvo conformado por los datos de 240 trabajadores colombianos. Los autores emplearon la técnica de ecuaciones estructurales. Los resultados demuestran que la variable flow como mediadora entre la percepción del clima social y la descripción del trabajo, logran explicar el 25% de la variabilidad de la satisfacción y el bienestar en el trabajo.

En relación a estudios nacionales podemos mencionar a:

Angulo-Paredes *et al.* (2021), quienes tuvieron como propósito identificar la dimensión predominante del aprendizaje organizacional que influye en el bienestar laboral de los docentes de educación secundaria de Lima. La técnica empleada fue la regresión logística. El conjunto de datos estuvo conformado por una muestra de 200 maestros elegidos por muestreo probabilístico aleatorio simple. Los resultados demuestran que la técnica de regresión logística obtuvo un R^2 de Nagelkerke de 0.727 en la predicción de la satisfacción laboral. Los autores concluyen que el aprendizaje organizacional influye en el bienestar laboral de los docentes; así mismo determinaron que la disposición al trabajo, disponibilidad de recursos y el reconocimiento por su labor influyeron en la satisfacción laboral.

CAPÍTULO II
PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

Estudiar los distintos aspectos que estén relacionados con el docente continúa siendo importante. Uno de estos es la satisfacción laboral que se encuentra dentro de la dimensión emocional (Hassan y Ibourk, 2021). Es importante mencionar que en el ámbito educativo se han encontrado las tasas más altas de estrés, ansiedad e irritabilidad que afectan el nivel de satisfacción laboral de los docentes (Serrano-García *et al.*, 2015).

La satisfacción laboral del docente es considerada un factor clave en la efectividad no solo del mismo, sino también de los estudiantes, el entorno escolar y el sistema educativo en general (Lopes y Oliveira, 2020; Sadeghi *et al.*, 2021). Actualmente se sabe que la satisfacción o insatisfacción de los docentes con su vida laboral podría tener impactos determinantes en la calidad de la enseñanza y el rendimiento de los estudiantes (Aouadni y Rebai, 2016; Lee y Nie, 2014). En el Perú, de acuerdo con Araoz-Estrada y Ramos-Gallegos (2021), el 42.1% de docentes presenta un nivel de satisfacción regular, es decir tienen una actitud poco positiva sobre su trabajo. En este escenario, es de suma importancia realizar estudios para entender de mejor manera la satisfacción laboral de los profesionales de la enseñanza, con métodos y herramientas actuales. Acorde con esta idea Ruiz-Quiles *et al.* (2015) recomiendan realizar estudios que permitan el análisis de cómo evitar la falta de satisfacción, para así evitar síntomas de agotamiento y desmotivación en los docentes.

Estudios próximos a esta problemática vienen empleando técnicas de aprendizaje automático como la regresión logística, bosques aleatorios, árboles de decisión, algoritmo AdaBoost, K-Vecinos más cercanos, máquinas de soporte vectorial y Naive Bayes, con gran éxito en la predicción de la deserción de empleados en varias

organizaciones (Gabrani y Kwatra, 2018; Sisodia *et al.*, 2018; Yogesh *et al.*, 2020). Estas técnicas, también están siendo empleadas para explorar y descubrir predictores confiables en grandes volúmenes de datos con gran número de dimensiones (Homocianu *et al.*, 2020). Asimismo, es posible encontrar factores que afectan la satisfacción laboral de los empleados mediante los algoritmos de *Gradient boosting* y Bosques aleatorios (Saisanthiya *et al.*, 2020). Otro estudio desarrollado por Seok *et al.* (2021), demuestra el uso de las redes neuronales artificiales para encontrar la asociación entre las variables demográficas, el liderazgo de coaching y la satisfacción laboral de los profesores de escuelas primarias en Corea del Sur con mejores resultados que la regresión múltiple.

La predicción de la satisfacción laboral de los empleados también se abordó mediante el empleo de redes neuronales profundas (Rustam *et al.*, 2021), algoritmo de árboles de decisión (CART) y redes neuronales (Tao Chen et al., 2021). Si bien, existen importantes avances en relación a esta temática, la mayoría de ellos abordan la satisfacción laboral de manera más general, es decir, no están enfocados a la satisfacción laboral del docente. Por otra parte, existen estudios que intentan explicar la satisfacción laboral del docente haciendo uso de técnicas como los modelos jerárquicos lineales (Hong-Hua *et al.*, 2016), modelo de ecuaciones estructurales (Tomás *et al.*, 2019) y análisis de varianza (ANOVA) (Asadujjaman *et al*., 2020).

2.2. Enunciado del problema

Identificado el contexto del problema, nos formulamos la siguiente interrogante general:

- ¿Cuál de los modelos predictivos machine learning de regresión logística, árboles de decisión y clasificadores combinados permitirá obtener los mejores indicadores en la predicción de la satisfacción laboral de docentes de educación básica del Perú?

Y las siguientes interrogantes específicas:

- ¿Cuáles son los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica?
- ¿Cuál es el modelo de regresión logística para la satisfacción laboral de docentes de educación básica del Perú?

- ¿Cómo es la predicción utilizando árboles de decisión para la satisfacción laboral de docentes de educación básica del Perú?

- ¿Cómo es la predicción utilizando clasificadores combinados para la satisfacción laboral de docentes de educación básica del Perú?

- ¿Cuáles es el resultado de comparar la técnica de regresión logística respeto a las técnicas de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú?

2.3. Justificación

La importancia del presente estudio está fundamentada en la necesidad de encontrar modelos predictivos machine learning para la satisfacción laboral de docentes de educación básica regular del Perú, esto debido a que la satisfacción laboral es uno de los principales factores que afectan el rendimiento de los estudiantes, así como la retención de estudiantes y maestros (Yoo y Rho, 2020). Adicional a los modelos, el presente estudio hará posible identificar predictores confiables para la satisfacción e insatisfacción laboral de los docentes desde un enfoque de aprendizaje automático, permitiendo de esta manera dotar de información a los funcionarios del sector educación que ayude en una toma de decisiones acertadas en políticas educativas, con el objetivo de garantizar una educación de calidad que conlleve al desarrollo del país (Asadujjaman et al., 2020; Homocianu et al., 2020). Los resultados también servirán como base para futuras investigaciones de la aplicación de algoritmos de inteligencia artificial para la predicción de la satisfacción laboral de los profesionales de la enseñanza, que de acuerdo con la revisión de la literatura es un campo con insipientes aplicaciones del aprendizaje automático. Para el logro de los objetivos planteados, será necesario el empleo metodologías, técnicas y herramientas que provienen de distintos campos de estudio como: la minería de datos y el aprendizaje automático, mismos que son de interés de profundización de conocimientos del investigador.

2.4. Objetivos

2.4.1. Objetivo general

Determinar cuál de los modelos predictivos *machine learning* de regresión logística, árboles de decisión y clasificadores combinados permitirá obtener los mejores

indicadores en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

2.4.2. Objetivos específicos

- Identificar los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica.
- Establecer el modelo logístico para la satisfacción laboral de docentes de educación básica del Perú.
- Predecir utilizando árboles de decisión la satisfacción laboral de docentes de educación básica del Perú.
- Predecir utilizando clasificadores combinados la satisfacción laboral de docentes de educación básica del Perú.
- Comparar las métricas de la técnica de regresión logística respecto a la técnica de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

2.5. Hipótesis

2.5.1. Hipótesis general

Los modelos predictivos *machine learning* de regresión logística, árboles de decisión y clasificadores combinados obtendrán una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

2.5.2. Hipótesis específicas

- El modelo construido con la técnica de regresión logística obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.
- El modelo construido con la técnica de árboles de decisión obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

- El modelo construido con la técnica de clasificadores combinados Random Forest obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. Lugar de estudio

Para el presente estudio se consideró la información de los docentes de todo el Perú, procedentes de la encuesta nacional a docentes de instituciones públicas de educación básica regular del Perú del año 2018, por ello podemos afirmar que el lugar de estudio corresponde a la nación del Perú; sin embargo, los experimentos tendrán lugar en la ciudad de Puerto Maldonado.

3.2. Población

De acuerdo con la ficha técnica de la encuesta nacional a docentes de instituciones educativas públicas y privadas - ENDO 2018, la cantidad de docentes que laboraron en el sector público y privado en el año 2018 asciende a 470931, lo que constituye nuestra población.

3.3. Muestra

De acuerdo con las especificaciones de la unidad de estadística del MINEDU, para el cálculo del tamaño de la muestra se utilizó la fórmula de muestreo aleatorio estratificado con afijación proporcional y considerando la varianza en cada estrato a partir de los datos de la ENDO 2016. Lo que permitirá obtener estimaciones con una confianza del 95% y con precisiones menores al 5.5%, según el nivel de inferencia estudiado. Tras este procedimiento se obtuvo una muestra a nivel global de 15,092 entrevistas (MUNEDU, 2018).

3.4. Método de investigación

El presente estudio se enmarca en la ciencia aplicada, de acuerdo con Bunge (1999) la investigación aplicada "es el campo de investigación en el que los problemas científicos con un posible sentido práctico se investigan con base en los descubrimientos de la ciencia básica (pura). Más que ser una investigación libre, tiene un objetivo, o mandato: de esa investigación se esperan eventualmente descubrimientos de interés práctico" (p. 277).

El fundamento epistemológico de este tipo de estudios son las distinciones tales como "Saber y Hacer", "Verdad y Acción", "Know-what y Know-How", "Conocimiento y Práctica", "Explicación y Aplicación", "Verdad y Eficiencia", etc. La idea de fondo está en las relaciones de utilidad del conocimiento, considerando que la función elemental del conocimiento en los organismos va estrechamente asociada a sus necesidades de subsistencia mediante mecanismos de adaptación al medio y control del mismo (Padrón, 2016).

3.5. Descripción detallada de métodos por objetivos específicos

a) Descripción de variables analizadas en los objetivos específicos

Modelo predictivo *Machine Learning.* Se refiere a modelos construidos a partir de conjuntos de grandes conjuntos de datos (Camacho et al., 2018). Lary *et al.* (2016) lo denominan aproximadores universales que aprenden del comportamiento subyacente de un sistema a partir de un conjunto de datos de entrenamiento. La siguiente expresión denota el paradigma general de un modelo predictivo:

$$Objetivo + Muestra + Algoritmo = Modelo\ predictivo$$

Donde el objetivo representa el problema a resolver. La muestra es el subconjunto representativo de la población. El algoritmo se refiere a algoritmos de aprendizaje automático y algoritmos de optimización del modelo. La combinación de estas da como resultado un modelo predictivo (Liu *et al.*, 2017).

Igual y Seguí (2020), afirman que el aprendizaje automático puede considerarse un subcampo de la inteligencia artificial y se divide en: Aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. En relación a los algoritmos para el aprendizaje supervisados, los autores mencionan como ejemplo: la regresión

logística, máquinas de soporte vectorial, árboles de decisión y bosques aleatorios (p. 67). Dentro de esta categoría también se encuentran las redes neuronales (Beunza-Nuin et al., 2020), *Gradiente Boosting* y *Extreme Gradient Boosting* (XGBoost) (Espinosa-Zúñiga y Espinosa-Zúñiga, 2020) .

Satisfacción laboral de docentes. La satisfacción laboral de acuerdo con Chiang-Vega *et al.* (2018) hace referencia al estado emocional positivo o placentero como resultado de la percepción subjetiva de las experiencias laborales de la persona. Continúa aclarando que este es un concepto global que refiere a las actitudes de las personas hacia diversos aspectos en su trabajo. En relación a la satisfacción del docente, no existe una definición generalmente acordada (Zembylas y Papanastasiou, 2004). Un intento de aproximación a esta definición la dan Skaalvik y Skaalvik (2011), cuando afirman que la satisfacción laboral del docente se refiere a las reacciones afectiva de los docentes a su trabajo o a su función docente.

En la Encuesta Nacional a Docentes (ENDO-2018), se encuentran aspectos como: Trabajo en sí mismo, Reconocimiento, Compañeros de trabajo, Pago, Condiciones de trabajo y Seguridad, que son aspectos de la satisfacción laboral docente considerados en el Cuestionario de Satisfacción Laboral para Profesores (TJSQ) originalmente elaborado por Lester en 1987, que consta de 66 reactivos en 9 factores: Supervisión, Compañeros de trabajo, Condiciones de trabajo, Pago, Responsabilidad, Trabajo en sí mismo, Progreso, Seguridad y Reconocimiento (Serrano-García *et al.*, 2015).

Para el presente estudio de predicción de la satisfacción laboral docente, se considerará como variable objetivo o target el trabajo docente en sí mismo.

b) Descripción detallada del uso de materiales, equipos, instrumentos, insumos, entre otros

La metodología empleada se muestra en la Figura 5, donde se observa las fases de: Limpieza y preprocesamiento de datos, selección de características y modelado predictivo. A continuación, procedemos a detallar cada uno de estas fases.

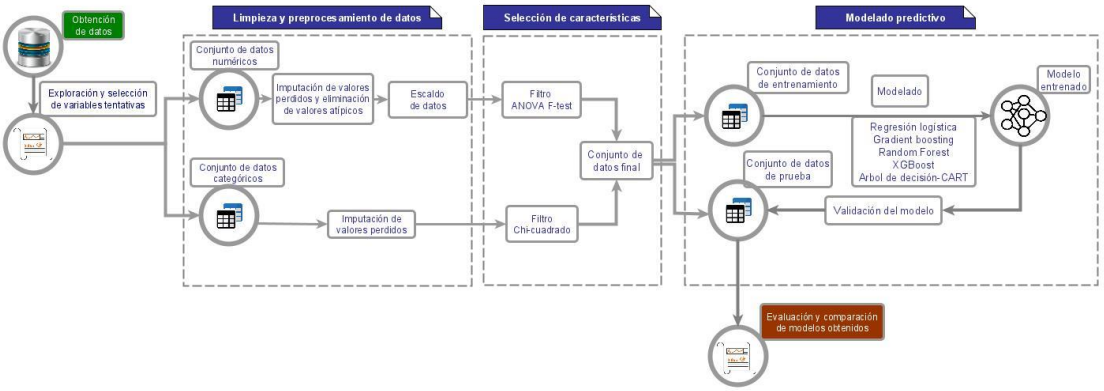


Figura 5. Metodología propuesta para el modelamiento de la satisfacción laboral de docentes

b.1. Limpieza y preprocesamiento de datos

El conjunto de datos utilizado para este estudio, fue obtenido de la Encuesta Nacional a Docentes (ENDO-2018), desarrollada por el Ministerio de Educación del Perú durante el año 2018 (MINEDU, 2022). La dimensión de este conjunto de datos es de 15087 instancias y 942 atributos. La información contenida en dicha encuesta, está divida en las siguientes categorías: Demografía, vivienda y hogar; Formación inicial; Trayectoria profesional y trabajo actual; Salud; Economía; Capacitación y formación en servicio; Medios de comunicación y tecnologías de la información; Opinión y percepción y Prácticas docentes. En la Tabla 2, podemos observar un resumen global del conjunto de datos generado por la función *create_report()* del paquete *DataExplorer* del lenguaje R.

Tabla 2

Resumen global del conjunto de datos ENDO-2018

Nombre	Valor
Filas	15087
Columnas	942
Columnas discretas	66
Columnas continuas	873
Todas las columnas que faltan	3
Observaciones faltantes	4717895
Filas completas	0
Observaciones totales	14211954
Asignación de memoria	111.8Mb

De las 942 columnas, se realizó una primera selección manual de 407 variables pertenecientes a las 9 categorías antes mencionadas. Se tomó como criterio de exclusión aquellas columnas que tienen un solo valor para todas las filas, así como las que poseían más de 40% de valores faltantes, debido a que estas pueden causar errores o resultados inesperados, como lo sugiere (Brownlee *et al.*, 2020). En este punto es importante aclarar que algunas variables estaban vacías por la naturaleza de las preguntas. Por ejemplo, la columna P323 que corresponde a la pregunta: Si pudiese volver hacia atrás, ¿elegiría nuevamente ser docente? con alternativas Sí, No; el hecho de responder afirmativamente esta pregunta, implicaba dejar vacía la columna P324 correspondiente a la pregunta: ¿Por qué no elegiría nuevamente ser docente?, para pasar a la pregunta 325. Por lo que fue de suma importancia el análisis y entendimiento de cada una de las variables. Tras este análisis se procede a filtrar columnas que poseen no más del 20% de datos perdidos, quedando así 370 columnas. Cabe aclarar que en esta encuesta existían valores con código NEP que correspondía a las celdas donde el encuestado no especificó la respuesta, por lo que también fue necesario imputar.

En relación a la variable objetivo, se construyó a partir de las respuestas a la pregunta: Considerando todas las cosas, ¿Cuán satisfecho está usted con su empleo en esta institución educativa?, la misma que se encuentra en la categoría Opinión y percepción de la encuesta en cuestión.

Imputación de valores perdidos y eliminación de valores atípicos

En esta tarea se procede a separar el conjunto de datos en numéricas con 40 columnas y categóricas 330 columnas, luego se imputó por moda para las variables categóricas y por la mediana las variables numéricas. De acuerdo con Navarro-Pastor y Losilla-Vidal (2000) y Useche y Mesa (2006) este procedimiento consiste en llenar el dato faltante de cada variable por la moda cuando se trata de una variable categórica y por la mediana para las variables numéricas. Es necesario aclarar que el conjunto de datos numéricos presentaba valores extremos, razón por la que se optó por imputar los valores faltantes con la mediana, debido que esta medida es más representativa y no es sensible a valores extremos. A este tipo de imputaciones en el ámbito estadístico se denominan métodos de imputación simple (Cuesta *et al.*, 2013; Rosati, 2017).

El algoritmo denominado Local Outlier Factor (LOF) permitió identificar 1785 instancias como valores extremos en el conjunto de datos numéricos. Tras la identificación, dichas instancias fueron removidas paralelamente en ambos conjuntos de datos. Para este algoritmo, el grado en el que un objeto es extremo depende de la densidad de su vecindad local, y es posible asignar un valor de LOF a cada objeto que representaría la probabilidad de que este sea un valor extremo (Alshawabkeh *et al.*, 2010). Para esta técnica los valores atípicos se identifican como puntos con densidad considerablemente más baja en comparación a sus vecinos.

Escalado de datos robusto

Este método de estandarización se aplicó debido a que trabaja con la mediana en lugar de la media, como se sabe la mediana es más representativa cuando existen valores atípicos. Para ello se calcula la mediana (percentil 50) y los percentiles 25 y 75. Luego, se procede a restar la mediana a los valores de cada variable y dividimos entre el valor del rango intercuartílico IQR=p_{75}-p_{25}. Los nuevos valores tienen una media y mediana de cero y una desviación estándar de 1 (Brownlee *et al.*, 2020). La fórmula empleada para este tipo de escalado es la siguiente:

$$valor\_escalado = \frac{valor - mediana}{p_{75} - p_{25}}$$

Donde: p_{75} y p_{25} son los percentiles 75 y 25 respectivamente.

La implementación de este método de escalado se encuentra en la clase *RobustScaler* del módulo *preprocessing* de *sklearn*.

b.2. Selección de características

En esta fase se empleó la clase *SelectKBest* de la librería de aprendizaje automático *Scikit-learn*. Esta implementación permitió la selección de las mejores características basadas en pruebas estadísticas bivariados. De acuerdo con Pedregosa *et al.* (2011) esta clase permite seleccionar las *k* características que tengan las puntuaciones más altas. Para determinar estas puntuaciones, cuando las variables de entrada sean numéricas y variable objetivo categórico es posible usar el estadístico F de ANOVA implementado en la función *f_casific()* y cuando ambas son categóricos podemos usar el estadístico Chi-cuadrado implementado en la función *chi2()* de la librería en referencia (Brownlee *et al.*, 2020). Se emplea los métodos de filtro debido a que las características se seleccionan en función de medidas estadísticas. Este procedimiento es independiente del algoritmo de aprendizaje y requiere menos tiempo computacional. Por otro lado los métodos de envoltorio son computacionalmente más costosos que los métodos de filtro, debido a los pasos de aprendizaje repetidos y validación cruzada (Chandrashekar y Sahin, 2014; Khaire y Dhanalakshmi, 2022; Miao y Niu, 2016). Otro enfoque para la selección de características es el método embebido son computacionalmente menos intensivos, sin embargo, este método tiene el inconveniente de ser específico de un modelo de aprendizaje (Venkatesh y Anuradha, 2019).

Filtro ANOVA F-test

Para la selección de características mediante este filtro se estableció como parámetros *score_func=f_classif* y *k='all'* en la clase *SelectKBest*, dando como entrada el conjunto de datos numérico y la satisfacción laboral del docente como variable objetivo.

La ecuación que permite obtener estas puntuaciones corresponde al estadístico F, que está dado por:

$$F = \frac{\frac{SSBG}{df_1}}{\frac{SSWG}{df_2}}$$

Donde:

SSBG=Suma de cuadrados entre grupos:

$$SSBG = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

SSWG=Suma de cuadrados dentro de los grupos:

$$SSWG = \sum_{j=1}^{m} n_j(\bar{y}_j - \bar{y})^2$$

Además, $df_1 = m - 1, df_2 = n - m, m$ es el número de niveles de nuestra variable objetivo y *n* es el número de observaciones, $\bar{y}_j$ es la media en el grupo $j, \bar{y}$ es la media total (Montgomery, 2004).

Filtro Chi-Cuadrado

Para usar este filtro se empleó la clase *SelectKBest* de la librería antes mencionada, con los parámetros *score_func=chi2* y *k='all'*. Para utilizar este filtro primero se realizó la codificación ordinal de los atributos categóricos (Dashdondov *et al.*, 2021). Esta codificación se detalla líneas más abajo. La puntuación, devuelta por la función mencionada, es calculada mediante el estadístico Chi-cuadrado de Pearson (Quintero y Duran, 2004). A continuación, se muestra dicha fórmula:

$$\bar{X}^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$

Donde:

O_i son los valores observados

E_i son los valores esperados

El estadístico Chi-cuadrado de Pearson es una prueba de independencia entre dos variables categóricas (Brownlee et al., 2020).

Construcción del conjunto de datos final

Para esta tarea se procede a codificar las variables nominales con el módulo *sklearn.preprocessing.OneHotEncoder* y las variables ordinales con el módulo *sklearn.preprocessing.OrdinalEncoder* como lo sugiere el manual oficial de Sklearn (Pedregosa *et al.*, 2011) y el estudio presentado por (Hancock y Khoshgoftaar, 2020).

La codificación *OneHotEncoder* transforma una variable con *n* observaciones y *k* categorías en k variables binarias con n observaciones cada una (Potdar, 2017; Zheng y Casari, 2018). Por su parte la codificación *OrdinalEncoder*, asigna un entero a cada categoría, siempre que se conozca el número de categorías existentes. En el caso de la codificación ordinal, no se agrega ninguna columna nueva a los datos; sin embargo, implica un orden para la variable (Potdar, 2017).

b.3. Modelado predictivo

Conjunto de datos de entrenamiento y validación

En esta tarea se realizó la partición del conjunto de datos en 70% para el entrenamiento y 30% para la validación de los modelos, como lo realizado por (Fallucchi et al., 2020; Moon et al., 2021a). Los resultados de esta partición nos dan 9911 instancias para entrenar y 3991 instancias para validar.

Tras la partición del conjunto de datos, se procede a verificar la distribución de las clases en la variable objetivo del conjunto de datos de entrenamiento. Esta distribución se muestra en la Figura 6 ítem a, donde podemos observar que esta variable se encuentra parcialmente desbalanceada. De acuerdo a Torres-Vásquez *et al.* (2021) y Bourel *et al.* (2021), el desbalanceo de datos puede afectar el resultado de los modelos de clasificación, por cuanto tienden a sesgar los resultados hacia la clase mayoritaria. Este problema se resolvió realizando el balanceo de datos al conjunto de datos de entrenamiento, empleando la técnica de sobremuestreo que consiste en añadir datos a la clase minoritaria hasta que se logre un equilibrio en ambas clases. En la Figura 6 ítem b podemos observar la distribución de las clases luego del balanceo de datos.

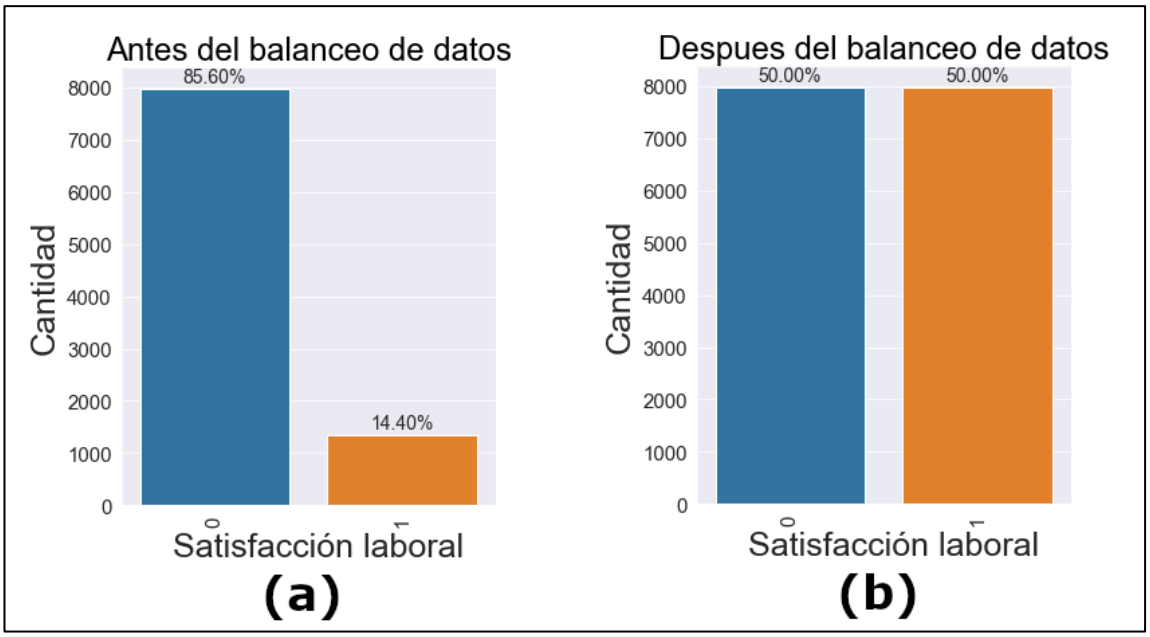


Figura 6. Distribución de clases en la variable objetivo.

Modelado

En esta tarea se empleó los algoritmos de Regresión logística, *Gradient Boosting*, *Random Forest*, *XGBoost y Decision Trees-CART* debido a que estos algoritmos se vienen empleando en la predicción de la satisfacción laboral con buenos desempeños, tal como se evidencia en los estudios presentados por (Arambepola y Munasinghe, 2021b; Tao Chen et al., 2021; Pratt et al., 2021; Rustam et al., 2021; Saleh y Abu-Soud, 2021). Para el entrenamiento de los modelos se consideraron los valores de hiperparámetros por defecto que la librería scikit-learn 1.0.1 establece, a excepción de parámetro *max_depth*=6 debido a que valores altos de este podrían ocasionar un sobreajuste del modelo; es decir, que el modelo tenga un buen desempeño con los datos entrenados pero pésimos resultados con datos nuevos. A continuación, se muestra la configuración de hiperparámetros:

Regresión logistica: Valores por defecto.

Gradient boosting: {'max_depth': 6}

Random Forest: {'max_depth': 6}

XGBoost: {'max_depth': 6}

Decision Trees-CART: {'max_depth': 6, criterion='gini'}

Validación de modelos

La validación de los modelos se realizó en el conjunto de datos de prueba. Los resultados del presente estudio se obtuvieron de esta tarea.

Evaluación y comparación de modelos obtenidos

En este punto se tuvo la necesidad de calcular las métricas de exactitud, sensibilidad, especificidad, área bajo la curva ROC (AUC), valor predictivo positivo, valor predictivo negativo, sin embargo, debido a que la distribución de las clases en la variable objetivo no se encuentra equilibrada se selecciona el mejor modelo en función a la sensibilidad, área bajo la curva ROC (AUC), menor número de falsos negativos y mayor número de verdaderos positivos que se obtienen la matriz de confusión.

c) Aplicación de prueba estadística inferencial

Para esta tarea se emplearon las métricas de exactitud, sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y el área bajo la curva ROC (Gironés *et al.*, 2017).

Adicional a las métricas mencionadas se empleó la distribución t-student para el contraste de hipótesis:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Identificar los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica

Para la selección de las variables más influyentes, se realiza el análisis de la relación entre el número de características y la obtención de exactitud en un modelo de regresión logística, con una validación cruzada estratificada con 10 pliegues y 3 repeticiones. Este procedimiento primero se aplica al conjunto de datos categóricos con el filtro Chi-cuadrado (Figura 7); luego, en el conjunto de datos numéricos con el filtro ANOVA F-test (Figura 8).

```
from numpy import mean
from numpy import std
import pandas as pd
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline

X, y = X_train,Ytrain

def evaluate_model(model):
    cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
    scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
    return scores

num_features = [i+1 for i in range(X.shape[1])]

results = list()
accuracy=[]
for k in num_features:
    model = LogisticRegression(solver='liblinear')
    fs = SelectKBest(score_func=chi2, k=k)
    pipeline = Pipeline(steps=[('chi',fs), ('lr', model)])
    scores = evaluate_model(pipeline)
    results.append(scores)
    print('>%d %.6f (%.6f)' % (k, mean(scores), std(scores)))
    accuracy.append([k,mean(scores), std(scores)])
df_accuracy_cat = pd.DataFrame(accuracy, columns = ['number of feature','mean accuracy', 'std accuracy'])
df_accuracy_cat.to_excel("evolucion_exactitud_cat.xlsx")
```

Figura 7. Script para la obtención de la exactitud en variables categóricas

```python
from numpy import mean
from numpy import std
import pandas as pd
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline

X, y = X_train,Ytrain

def evaluate_model(model):
    cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
    scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
    return scores

num_features = [i+1 for i in range(X.shape[1])]

results = list()
accuracy=[]
for k in num_features:
    model = LogisticRegression(solver='liblinear')
    fs = SelectKBest(score_func=f_classif, k=k)
    pipeline = Pipeline(steps=[('anova',fs), ('lr', model)])
    scores = evaluate_model(pipeline)
    results.append(scores)
    print('>%d %.6f (%.6f)' % (k, mean(scores), std(scores)))
    accuracy.append([k,mean(scores), std(scores)])
df_accuracy_num = pd.DataFrame(accuracy, columns = ['number of feature','mean accuracy', 'std accuracy'])
df_accuracy_num.to_excel("evolucion_exactitud_num.xlsx")
```

Figura 8. Script para la obtención de la exactitud en variables numéricas

En la Figura 9, podemos observar que el margen de ganancia en exactitud para el modelo no es significativo a medida que aumenta el número de características, por ello se elige nueve características del conjunto de datos categóricos y uno del conjunto de datos numéricos.

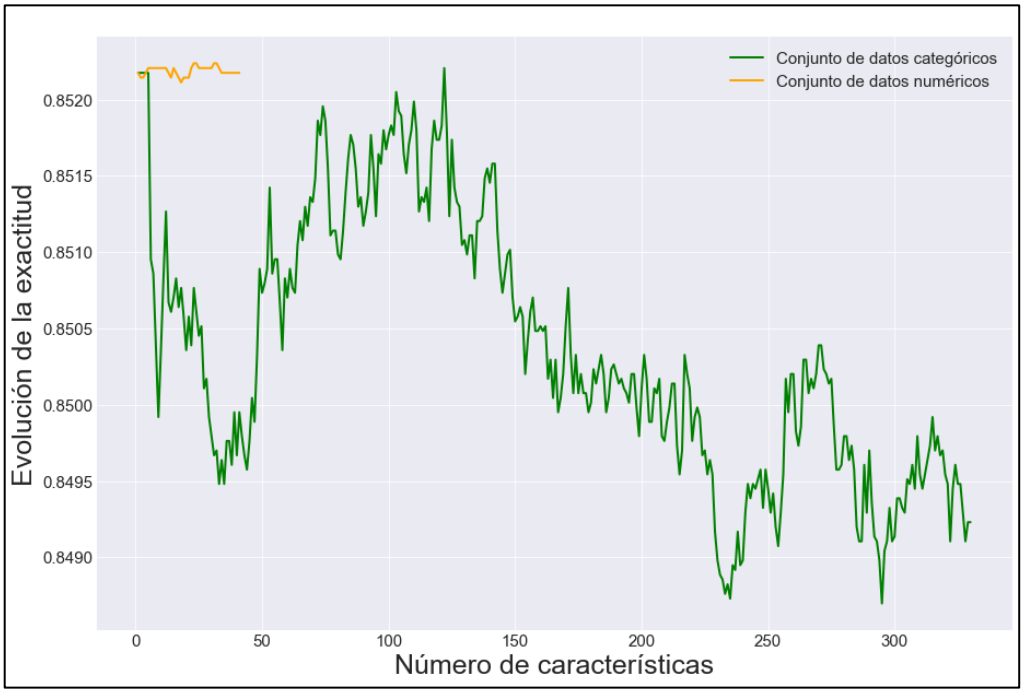


Figura 9. Evolución de la exactitud y número de características

Conocido el comportamiento de la exactitud en relación a la cantidad de características, se procede a obtener las puntuaciones Chi-Cuadrado y ANOVA de las variables predictoras sobre la variable objetivo. En la Figura 10 se observa el código fuente para la obtención de las puntuaciones Chi-Cuadrado.

```
from sklearn.preprocessing import OrdinalEncoder
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

def prepare_inputs(X_train):
    oe = OrdinalEncoder()
    oe.fit(X_train)
    X_train_enc = oe.transform(X_train)
    return X_train_enc

def select_features_chi2(X_train, Ytrain):
    fs = SelectKBest(score_func=chi2, k='all')
    fs.fit(X_train, Ytrain)
    X_train_fs = fs.transform(X_train)
    return X_train_fs,fs

X_train_enc= prepare_inputs(df_cat_tr)

X_train_fs,fs = select_features_chi2(X_train_enc,Ytrain)

feature=[]
for i in range(len(fs.scores_)):
    feature.append([df_cat_tr.columns[i],fs.scores_[i]])
df_feature = pd.DataFrame(feature, columns = ['Variable','Score'])

df_feature=df_feature.sort_values('Score',ascending=False).reset_index(drop=True)
df_feature.iloc[0:50].to_excel("var_imp_cat.xlsx")
df_feature
```

Figura 10. Script para la obtención de puntuaciones Chi-Cuadrado

La Figura 11 muestra las puntuaciones Chi-Cuadrado en orden de importancia de las 50 variables categóricas más influyentes en la predicción de la satisfacción laboral docente.

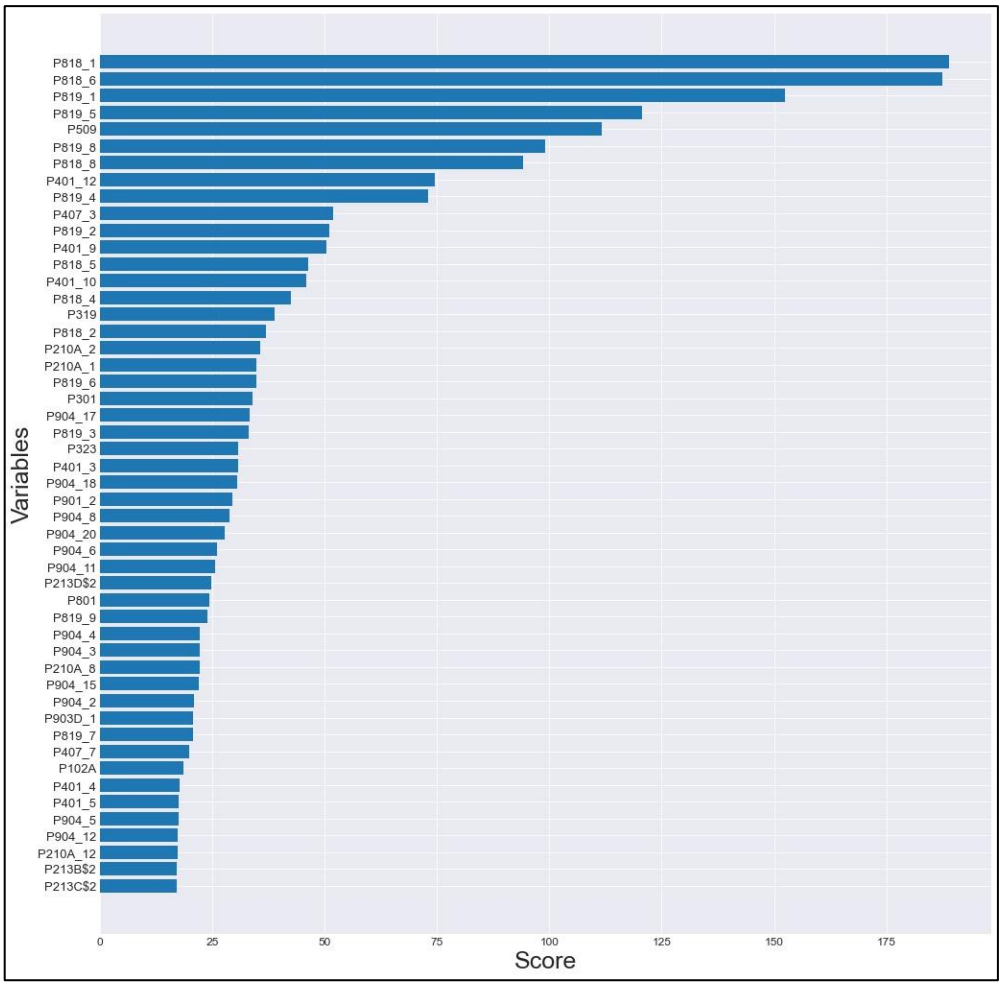


Figura 11. Puntuaciones Chi-Cuadrado en la predicción de la satisfacción laboral docente

En la Figura 12 se observa el código fuente para la obtención de las puntuaciones ANOVA.

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
from matplotlib import pyplot
def select_features_anova(X_train, y_train):
    fs = SelectKBest(score_func=f_classif, k='all')
    fs.fit(X_train, y_train)
    X_train_fs = fs.transform(X_train)
    return X_train_fs, fs

X_train_fs, fs = select_features_anova(df_num_tr, Ytrain)

feature=[]
for i in range(len(fs.scores_)):
    feature.append([df_num_tr.columns[i],fs.scores_[i]])
df_feature_num = pd.DataFrame(feature, columns = ['Variable','Score'])

df_feature_num=df_feature_num.sort_values('Score',ascending=False).reset_index(drop=True)
df_feature_num.to_excel("var_imp_num1.xlsx")
df_feature_num
```

Figura 12. Script para la obtención de puntuaciones ANOVA

La Figura 13 muestra las puntuaciones ANOVA en orden de importancia de las variables numéricas en la predicción de la satisfacción laboral docente.

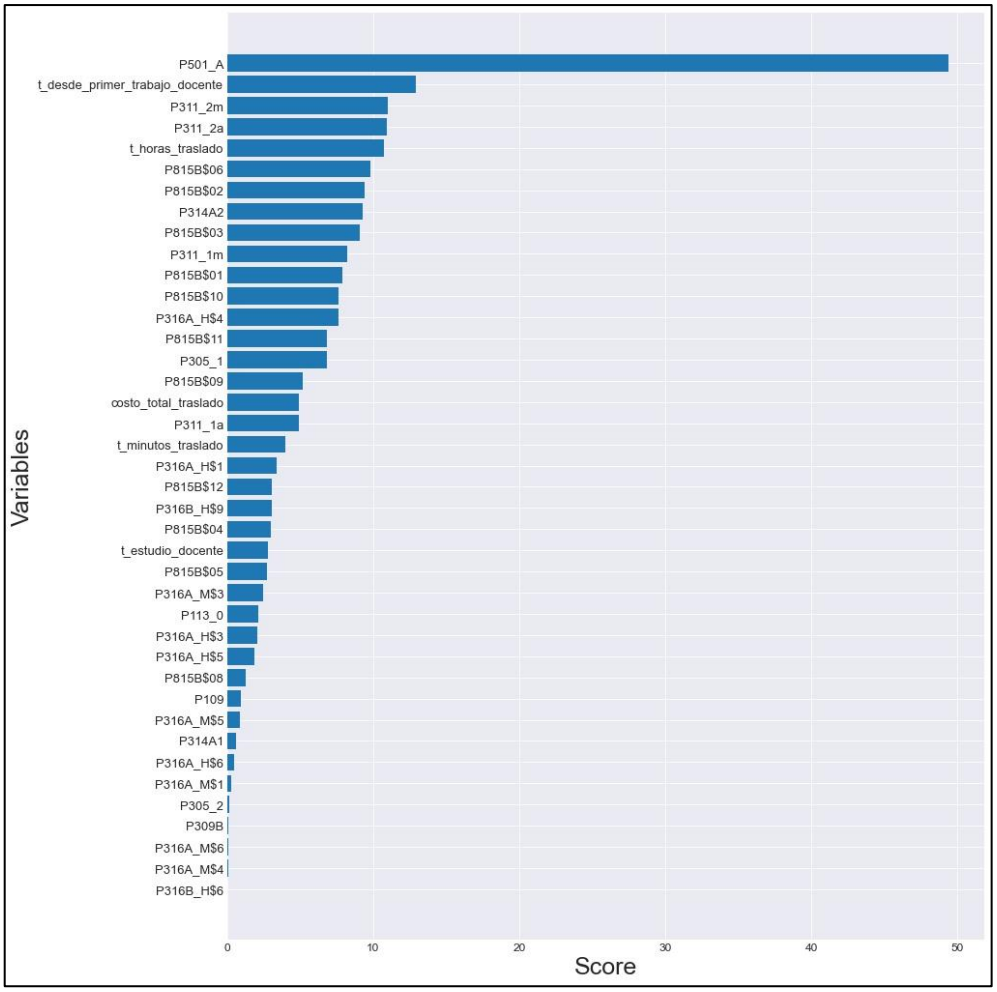


Figura 13. Puntuaciones ANOVA en la predicción de la satisfacción laboral docente

En la Tabla 3, se consolida la selección de las variables más influyentes en la predicción de la satisfacción laboral docente. La primera columna muestra la descripción de la variable y la segunda columna muestra la denominación original en el conjunto de datos ENDO-2018. A partir de esta tabla, se construye el conjunto de datos final para la predicción de la satisfacción laboral de docentes de educación básica.

Tabla 3

Descripción de las características seleccionadas

Descripción	Denominación
Satisfacción con su vida	P818_1
Satisfacción con su autoestima	P818_6
Satisfacción con su actividad pedagógica	P819_1
Satisfacción con la relación con el director (a)	P819_5
Percepción de las condiciones de vida	P509
Satisfacción con su salario	P819_8
Satisfacción con sus relaciones familiares	P818_8
El año pasado sufrió depresión	P401_12
Satisfacción de la relación con sus colegas	P819_4
Ingreso total (bruto)	P501_A

De esta tabla (Tabla 3), la variable P401_12 es de tipo nominal, P501_A es de tipo numérico continuo y el resto son de tipo ordinal.

Tras la selección de características más influyentes para el problema de predicción, se construye el conjunto de datos final.

La Figura 14 muestra la estructura del conjunto de datos final. La variable *Job Satisfaction* constituye nuestra variable objetivo, donde uno representa insatisfecho y cero representa satisfecho.

	P401_12_No	P401_12_Si	P501_A	P509	P818_1	P818_6	P818_8	P819_1	P819_4	P819_5	P819_8	JobSatisfaction
0	1	0	-0.065206	2	2	3	2	2	3	2	2	0
1	1	0	0.596419	2	2	3	3	2	2	2	3	0
2	1	0	-0.241081	2	3	3	3	3	3	3	3	0
3	1	0	-0.380665	2	1	2	2	2	3	3	1	1
4	1	0	-0.241081	2	2	3	3	3	1	2	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
13297	1	0	0.317252	1	2	2	2	2	2	2	1	0
13298	1	0	0.317252	2	2	2	2	2	2	2	1	1
13299	1	0	0.596419	2	2	2	2	2	2	2	2	0
13300	1	0	1.015169	2	3	2	2	2	3	2	3	0
13301	1	0	1.015169	2	3	2	2	2	2	2	3	0

13302 rows × 12 columns

Figura 14. Conjunto de datos final

4.2. Establecer el modelo logístico para la satisfacción laboral de docentes de educación básica del Perú

Se construye el modelo logístico empleando la clase *sklearn.linear_model.LogisticRegression.* La Figura 15 muestra el código fuente para esta tarea.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_sm,Y_sm)
score=model.score(X_sm,Y_sm)
print("Exactitud: "+str('{:.5f}'.format(score)))
```

```
Exactitud: 0.73676
```

Figura 15. Script para la construcción del modelo de regresión logística

La Figura 16 muestra las matrices de confusión del modelo logístico en el conjunto de datos de entrenamiento (Figura 16 ítem a) y en el conjunto de datos de prueba (Figura 16 ítem b). En el ítem b del conjunto de datos de prueba se observa que el modelo clasifica 189 docentes (4.74%) como satisfechos cuando en realidad estos se encuentran insatisfechos, este valor se representa en dicha matriz como FN que significa falsos negativos. El modelo clasifica 458 docentes (11.48%) como insatisfechos cuando en realidad eran docentes insatisfechos, este valor se representa en dicha matriz como VP que significa verdaderos positivos. Finalmente, este modelo clasifica a 2434 docentes (60.99%) como satisfechos cuando en realidad se encuentran satisfechos, valor que se representa como VN o verdaderos negativos.

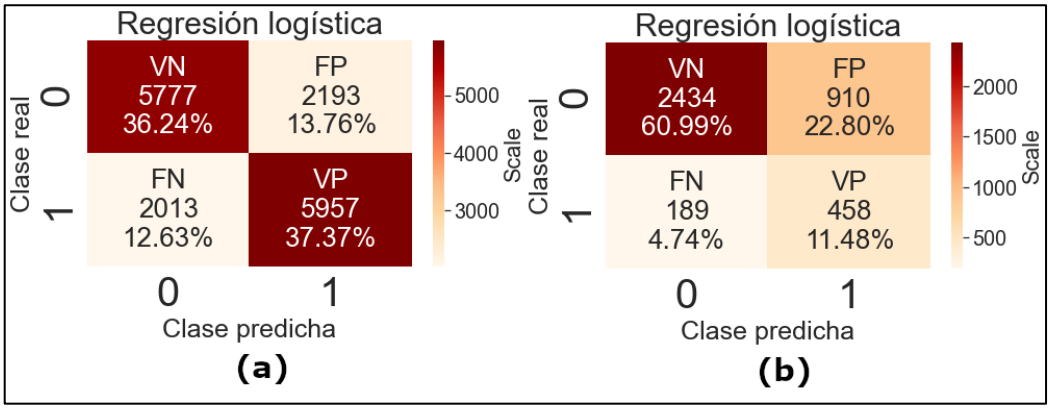


Figura 16. Matriz de confusión del modelo de regresión logística

A partir de esta matriz se calculan métricas de desempeño del modelo que se muestran en la Tabla 4.

Tabla 4

Métricas del modelo de regresión logística

Métrica	**Conjunto de datos de entrenamiento**	**Conjunto de datos de prueba**
Exactitud	0,73676	0,72463
Sensibilidad	0,74868	0,70788
Especificidad	0,72484	0,72787
Valor predictivo positivo	0,73125	0,33480
Valor predictivo negativo	0,74254	0,92795
F1-Score	0,73986	0,45459
AUC	0,81984	0,80194

La Tabla 4 muestra las métricas de desempeño del modelo de regresión logística en los conjuntos de datos de entrenamiento y prueba. La exactitud del modelo de regresión logística tiene un comportamiento similar en ambos conjuntos de datos, lo que indica que no existe sobreajuste del modelo, demostrando así que el modelo construido es robusto. Una Exactitud del 0.7246 indica una tasa de acierto en la predicción del 72.46% y una tasa de error en la predicción es del 27.54%.

La Tabla 4 también muestra una sensibilidad del 0.7079 (70.79%), esto representa la probabilidad de clasificar correctamente a un docente insatisfecho (Clase 1: Insatisfecho), respondiendo así a la pregunta: ¿Qué porcentaje de docentes insatisfechos han sido predichos por el modelo como insatisfechos?. La especificidad del modelo es del 0.7279 (72.79%) esto representa la probabilidad de clasificar correctamente a un docente satisfecho (Clase 0: Satisfecho), respondiendo así a la pregunta: ¿Qué porcentaje de docentes satisfechos han sido predichos por el modelo como satisfechos?. Los valores de las métricas de valor predictivo positivo, valor predictivo negativo y F1-Score se calcularon solo con fines de realizar la comparación de los modelos.

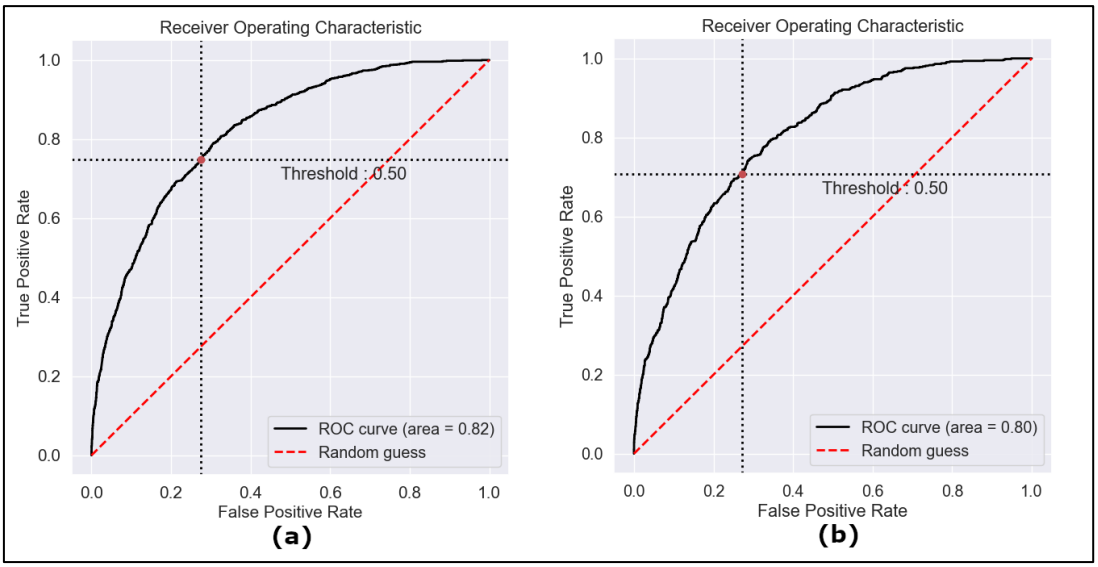


Figura 17. Área bajo la curva ROC (AUC) del modelo de regresión logística. (a) Gráfica del área bajo la curva (ROC) en el conjunto de datos de entrenamiento. (b) Gráfica del área bajo la curva (ROC) en el conjunto de datos de prueba.

En la figura 17 se observa las gráficas de las áreas bajo la curva ROC para el modelo de regresión logística en el conjunto de datos de entrenamiento (Figura 17 ítem a) y el conjunto de datos de prueba (Figura 17 ítem b), donde se observa que el valor del AUC en el conjunto de datos de prueba es de 0.80, esto demuestra que la capacidad discriminativa del modelo de regresión logística para las clases insatisfecho y satisfecho es bueno, afirmamos ello de acuerdo a la escala de valoración propuesta por (Gironés *et al.*, 2017)

Con el fin de establecer la ecuación del modelo de regresión logística, se procede a verificar la significancia estadística de las variables en la predicción de la satisfacción laboral de docentes.

La Figura 18 muestra los resultados de la significancia estadística de las variables predictoras, donde se observa que las variables 'P401_12_No', 'P401_12_Si' y 'P818_8' no resultaron ser significativos por cuanto el *p-value* para estas variables no cumplen la condición de *p-value*<0.05, por lo que se procede a ir retirando las variables de una a una para observar el desempeño del modelo resultante.

```
import statsmodels.api as sm

X_sm = sm.add_constant(X_sm, prepend=True)
modelo = sm.Logit(endog=Y_sm, exog=X_sm,)
modelo = modelo.fit()
print(modelo.summary())
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.516955
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:        JobSatisfaction   No. Observations:                15940
Model:                          Logit   Df Residuals:                    15929
Method:                           MLE   Df Model:                           10
Date:                Mon, 07 Feb 2022   Pseudo R-squ.:                  0.2542
Time:                        07:25:43   Log-Likelihood:                -8240.3
converged:                      False   LL-Null:                       -11049.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2360   3.69e+05   1.15e-05      1.000   -7.24e+05    7.24e+05
P401_12_No     1.9740   3.69e+05   5.35e-06      1.000   -7.24e+05    7.24e+05
P401_12_Si     2.2620   3.69e+05   6.13e-06      1.000   -7.24e+05    7.24e+05
P501_A        -0.4389      0.021    -20.894      0.000      -0.480      -0.398
P509          -0.3319      0.039     -8.414      0.000      -0.409      -0.255
P818_1        -0.5565      0.036    -15.335      0.000      -0.628      -0.485
P818_6        -0.1071      0.039     -2.743      0.006      -0.184      -0.031
P818_8        -0.0642      0.037     -1.759      0.079      -0.136       0.007
P819_1        -0.3517      0.042     -8.397      0.000      -0.434      -0.270
P819_4        -0.1734      0.039     -4.429      0.000      -0.250      -0.097
P819_5        -0.7111      0.034    -21.167      0.000      -0.777      -0.645
P819_8        -1.2850      0.032    -39.723      0.000      -1.348      -1.222
==============================================================================
```

Figura 18. Primera prueba de significancia estadística de variables predictoras

En la Figura 19, se muestra el resumen del modelo construido sin considerar las variables 'P401_12_Si', 'P818_8', tras ello se puede observar que todas las variables restantes resultaron ser significativos *p-value*<0.05.

```
X_sm_sig=X_sm[X_sm.columns.difference(['P401_12_Si','P818_8'])]

X_sm_sig = sm.add_constant(X_sm_sig, prepend=True)
modelo1 = sm.Logit(endog=Y_sm, exog=X_sm_sig,)
modelo1 = modelo1.fit()
print(modelo1.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.517052
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:        JobSatisfaction   No. Observations:                15940
Model:                          Logit   Df Residuals:                    15930
Method:                           MLE   Df Model:                            9
Date:                Mon, 07 Feb 2022   Pseudo R-squ.:                  0.2541
Time:                        07:35:25   Log-Likelihood:                -8241.8
converged:                       True   LL-Null:                       -11049.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
P401_12_No    -0.2944      0.065     -4.561      0.000      -0.421      -0.168
P501_A        -0.4393      0.021    -20.915      0.000      -0.480      -0.398
P509          -0.3349      0.039     -8.501      0.000      -0.412      -0.258
P818_1        -0.5642      0.036    -15.663      0.000      -0.635      -0.494
P818_6        -0.1267      0.037     -3.383      0.001      -0.200      -0.053
P819_1        -0.3634      0.041     -8.787      0.000      -0.444      -0.282
P819_4        -0.1809      0.039     -4.650      0.000      -0.257      -0.105
P819_5        -0.7120      0.034    -21.183      0.000      -0.778      -0.646
P819_8        -1.2855      0.032    -39.747      0.000      -1.349      -1.222
const          6.4717      0.146     44.331      0.000       6.186       6.758
==============================================================================
```

Figura 19. Segunda prueba de significancia estadística de variables predictoras

Tras haber obtenido las variables significativas, se establece la ecuación del modelo de regresión logística para predecir la probabilidad de que un docente se sienta insatisfecho con su labor (Clase positiva 1: Insatisfecho). Esta ecuación se muestra en la Figura 20.

$$p(y=1|X)=\frac{e^{(6.47-0.29*P401\_12\_No-0.44*P501\_A-0.33*P509-0.56*P818\_1-0.13*P818\_6-0.36*P819\_1-0.18*P819\_4-0.71*P819\_5-1.29*P819\_8)}}{1+e^{(6.47-0.29*P401\_12\_No-0.44*P501\_A-0.33*P509-0.56*P818\_1-0.13*P818\_6-0.36*P819\_1-0.18*P819\_4-0.71*P819\_5-1.29*P819\_8)}}$$

Figura 20. Modelo de regresión logística para la predicción de la satisfacción laboral docente

De acuerdo con la segunda prueba de significancia estadística de variables predictoras (Figura 19), podemos afirmar que este nuevo modelo en su conjunto es significativo (*Likelihood ratio p-value* = 0.000). Además, en la Figura 21 se observa que el porcentaje de clasificación correcto de este modelo en el conjunto de datos de prueba

es del 72.41% un valor muy por encima del umbral del 50% esperado por azar; por tanto, este nuevo modelo también es aceptable.

```
import statsmodels.api as sm
from sklearn.metrics import accuracy_score

Xtest_sig=Xtest[Xtest.columns.difference(['P401_12_Si','P818_8'])]

Xtest_sig = sm.add_constant(Xtest_sig, prepend=True)
Y_predict = modelo1.predict(exog = Xtest_sig)
clasificacion = np.where(Y_predict<0.5, 0, 1)
accuracy = accuracy_score(
            y_true    = Ytest,
            y_pred    = clasificacion,
            normalize = True
           )
print("Exactitud: "+str('{:.5f}'.format(accuracy)))
```

```
Exactitud: 0.72413
```

Figura 21. Exactitud del modelo de regresión logística

Análisis inferencial

Para este análisis se obtiene cien mediciones de las métricas más importantes en modelos predictivos, se optó por tomar 100 mediciones para cada uno de los modelos estudiados, los cuales son: Regresión logística, Árbol de decisión, Bosques aleatorios, *Gradient Boosting* y XGBoost, formando así un conjunto de datos de 500 registros que serán empleados para la contratación de hipótesis. Esta medición fue posible mediante la técnica de validación cruzada *K-Fold Cross-Validation* con *k=100*, que consiste en dividir el conjunto de datos en forma aleatoria en *k* grupos aproximadamente del igual tamaño, luego *k-1* grupos se emplean para el entrenamiento del modelo y un grupo para la validación del modelo. El proceso es iterativo y se repite *k* veces, empleando un grupo distinto para la validación en cada iteración (Amat-Rodrigo, 2016). La Figura 22 muestra las mediciones para la exactitud (columna 2), sensibilidad (columna 3), valor predictivo positivo (columna 4), puntuación F1 (columna 5) y el área bajo la curva ROC o AUC (columna 6) para todos los modelos.

	Model	Accuracy	Sensitivity	PPV	F1 Score	AUC
0	Logistic Regression	0.700000	0.7250	0.690476	0.707317	0.810000
1	Logistic Regression	0.756250	0.7500	0.759494	0.754717	0.826719
2	Logistic Regression	0.706250	0.7250	0.698795	0.711656	0.753047
3	Logistic Regression	0.668750	0.6000	0.695652	0.644295	0.787344
4	Logistic Regression	0.775000	0.7875	0.768293	0.777778	0.818984
...	...	...	...	...	...	...
495	Decision Trees-CART	0.798742	0.7500	0.833333	0.789474	0.881487
496	Decision Trees-CART	0.792453	0.7875	0.797468	0.792453	0.845411
497	Decision Trees-CART	0.811321	0.7875	0.828947	0.807692	0.882199
498	Decision Trees-CART	0.710692	0.7500	0.697674	0.722892	0.794541
499	Decision Trees-CART	0.729560	0.6750	0.760563	0.715232	0.820965

500 rows × 6 columns

Figura 22. Conjunto de datos de mediciones de métricas por modelo

La Tabla 5 muestra los valores AUC obtenidos con el modelo de Regresión logística, que serán empleados para la prueba de hipótesis específica 1.

Tabla 5

Valores AUC del modelo de Regresión logística

k	AUC	*k*	AUC	*k*	AUC	*k*	AUC	*k*	AUC
1	0,8100000	21	0,8103906	41	0,7871044	61	0,8264241	81	0,8455696
2	0,8267188	22	0,8234375	42	0,8409810	62	0,7329905	82	0,8122627
3	0,7530469	23	0,8407813	43	0,8863924	63	0,8355222	83	0,8206487
4	0,7873438	24	0,8881250	44	0,8554589	64	0,7884494	84	0,8314873
5	0,8189844	25	0,8285156	45	0,7985759	65	0,8604430	85	0,8246835
6	0,8057031	26	0,8145313	46	0,7933544	66	0,7892405	86	0,8049051
7	0,7993750	27	0,8282813	47	0,8371044	67	0,8917722	87	0,8145570
8	0,8040625	28	0,8254688	48	0,8022943	68	0,8345728	88	0,8146361
9	0,8491406	29	0,7658594	49	0,7820411	69	0,8381329	89	0,7946203
10	0,8151563	30	0,8456250	50	0,8126582	70	0,8626582	90	0,8222310
11	0,8441406	31	0,8056250	51	0,7887658	71	0,8151108	91	0,8772943
12	0,8503125	32	0,8268750	52	0,7994462	72	0,8011076	92	0,7911392
13	0,8225781	33	0,8459375	53	0,7492089	73	0,7426424	93	0,7974684
14	0,8700000	34	0,7932813	54	0,8416930	74	0,8015823	94	0,8665348
15	0,8753125	35	0,7926563	55	0,8096519	75	0,7738133	95	0,8159019
16	0,8015625	36	0,8103906	56	0,7885285	76	0,8359177	96	0,8242089
17	0,8581250	37	0,8095313	57	0,8580696	77	0,7973101	97	0,8284810
18	0,8270313	38	0,8313281	58	0,8468354	78	0,8066456	98	0,8613924
19	0,8504688	39	0,8057813	59	0,8636076	79	0,8045886	99	0,7855222
20	0,7994531	40	0,8394531	60	0,7943829	80	0,8183544	100	0,7864715

Prueba de normalidad

a) Formulación de hipótesis estadística

Hipótesis nula:

H_0: La variable AUC del modelo de regresión logística proviene de una distribución normal.

Hipótesis alterna:

H_1: La variable AUC del modelo de regresión logística no proviene de una distribución normal.

b) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

c) Elección del estadístico de contraste

Prueba de Lilliefors.

```
from statsmodels.stats.diagnostic import lilliefors
lilliefors_test_auc = lilliefors(df_lr.AUC)
lilliefors_test_auc
print("statistic: "+str('{:.5f}'.format(lilliefors_test_auc[0])))
print("pvalue: "+str('{:.5f}'.format(lilliefors_test_auc[1])))
```

```
statistic: 0.06861
pvalue: 0.29902
```

Figura 23. Resultados de prueba de normalidad del AUC del modelo de Regresión logística

d) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Como $p - valor = 0.29902 \geq 0.05$, no se rechaza H_0, por lo tanto, la variable de estudio AUC del modelo de regresión logística proviene de una distribución normal, para un nivel de confianza del 95%.

Contrastación de hipótesis

a) Hipótesis específica 1

El modelo construido con la técnica de regresión logística obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

b) Formulación de hipótesis estadística

Hipótesis nula:

H_0: $\mu \leq 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de regresión logística es menor o igual a 0.74.

Hipótesis alterna:

H_1: $\mu > 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de regresión logística es mayor a 0.74.

c) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

d) Elección del estadístico de contraste

Elegimos la prueba paramétrica de T-Student para una sola muestra, por cuanto el conjunto de datos sigue una distribución normal y se desconoce la desviación estándar poblacional. El cálculo del valor de este estadístico lo realizamos con la librería pcipy de Python. La Figura 24 muestra el valor de este estadístico.

```
from scipy import stats
stats.ttest_1samp(df_lr.AUC, popmean=0.75, alternative='greater')
```

```
Ttest_1sampResult(statistic=22.347621641818137, pvalue=9.176873792333366e-41)
```

Figura 24. Resultados de la prueba de hipótesis específica 1

e) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Dado que $p - valor$ =9.176873792333366e-41<0.05, entonces rechazamos H_0, y aceptamos H_1.

Existe evidencia estadística suficiente de que la media de las áreas bajo la curva ROC del modelo construido con la técnica de regresión logística es mayor a 0.74, con un nivel de confianza del 95%.

Por lo tanto, se comprueba la hipótesis específica 1, que indica: El modelo construido con la técnica de regresión logística obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

4.3. Predecir utilizando árboles de decisión la satisfacción laboral de docentes de educación básica del Perú

Para el logro de este objetivo se procede a crear el modelo de árbol de decisión, empleando la librería *sklearn.tree.DecisionTreeClassifier*. La figura 25 muestra el código fuente para realizar esta tarea.

```
from sklearn.tree import DecisionTreeClassifier
tree_cl = DecisionTreeClassifier(random_state=42, max_depth=6,criterion="gini")
tree_cl.fit(X_sm, Y_sm)
score=tree_cl.score(X_sm, Y_sm)
print("Exactitud: "+str('{:.5f}'.format(score)))
```

```
Exactitud: 0.75276
```

Figura 25. Script para la construcción del modelo de árbol de decisión

La Figura 26 muestra las matrices de confusión de árbol de decisión para el conjunto de datos de entrenamiento (Figura 26 ítem a) y en el conjunto de datos de prueba (Figura 26 ítem b). En el ítem b del conjunto de datos de prueba se observa que el modelo clasifica 250 docentes (6.26%) como satisfechos cuando en realidad estos se encuentran insatisfechos, este valor se representa en dicha matriz como FN que significa falsos negativos. Esta matriz también muestra que el modelo clasifica 397 docentes (9.95%) como insatisfechos cuando en realidad eran docentes insatisfechos, este valor se representa en dicha matriz como VP que significa verdaderos positivos. Esta matriz también logra clasificar a 2700 docentes (67.65%) como satisfechos cuando en realidad estos se encuentran satisfechos.

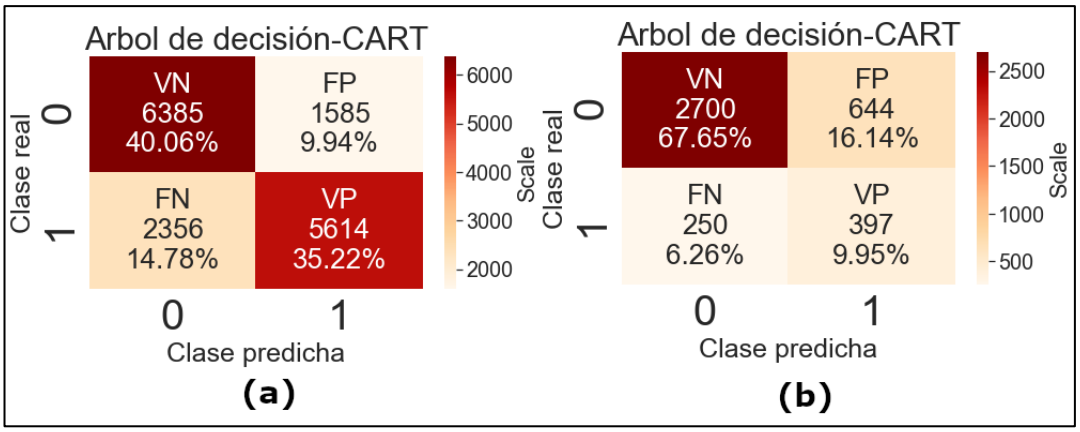


Figura 26. Matriz de confusión del modelo de árbol de decisión-CART

Con los valores de la matriz de confusión obtenidos en el conjunto de datos de prueba Figura 26, se procede a calcular las métricas del modelo construido con el algoritmo de árbol de decisión. Estos resultados se muestran en la Tabla 6.

Tabla 6

Métricas del modelo de árbol de decisión-CART

Métrica	**Conjunto de datos de entrenamiento**	**Conjunto de datos de prueba**
Exactitud	0,75276	0,77600
Sensibilidad	0,70439	0,61360
Especificidad	0,80113	0,80742
Valor predictivo positivo	0,77983	0,38136
Valor predictivo negativo	0,73047	0,91525
F1 Score	0,74019	0,47038
AUC	0,83527	0,80519

La Tabla 6 muestra las métricas de desempeño del modelo de árbol de decisión en los conjuntos de datos de entrenamiento y prueba. La tabla muestra que este modelo obtuvo una Exactitud del 0.776 en el conjunto de datos de prueba, lo que indica una tasa de acierto en la predicción del 77.6% y una tasa de error en la predicción es del 22.4%. En esta tabla (columna 3) también muestra una sensibilidad del 0.6136 (61.36%), esta medida representa la probabilidad de clasificar correctamente a un docente insatisfecho (Clase 1: Insatisfecho), respondiendo así a la pregunta: ¿Qué porcentaje de docentes insatisfechos han sido predichos por el modelo como insatisfechos?. La especificidad del modelo es del 0.8074 (80.74%) esto representa la probabilidad de clasificar correctamente a un docente satisfecho (Clase 0: Satisfecho), respondiendo así a la pregunta: ¿Qué porcentaje de docentes satisfechos han sido predichos por el modelo como satisfechos?. De estas dos métricas podemos afirmar que el modelo de árbol de decisión identifica mejor la clase 0 es decir docentes satisfechos. Los valores de las métricas de valor predictivo positivo, valor predictivo negativo y F1-Score se calcularon solo con fines de realizar la comparación de los modelos.

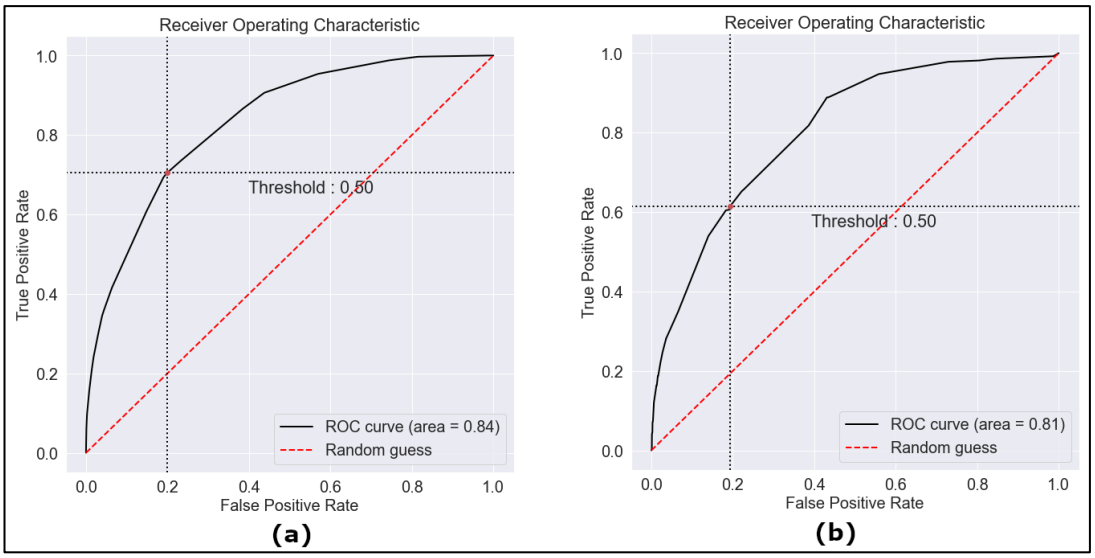


Figura 27. Área bajo la curva ROC (AUC) del modelo de árbol de decisión. (a) Gráfica del área bajo la curva (ROC) en el conjunto de datos de entrenamiento. (b) Gráfica del área bajo la curva (ROC) en el conjunto de datos de prueba.

En la figura 27, se observa las gráficas del área bajo la curva ROC del modelo de Árbol de decisión para el conjunto de datos de entrenamiento (Figura 27 ítem a) y el conjunto de datos de prueba (Figura 27 ítem b), donde observamos que el valor del AUC en el conjunto de datos de prueba es de 0.81. Esto demuestra que la capacidad discriminativa del modelo de árbol de decisión para las clases insatisfecho y satisfecho es bueno, afirmamos ello de acuerdo a la escala de valoración propuesta por (Gironés *et al.*, 2017).

Análisis inferencial

Se procede a realizar el análisis inferencial con los valores AUC obtenidos con el modelo de Árbol de decisión que se muestran en la Tabla 7.

Tabla 7

Valores AUC del modelo de Árbol de decisión

k	AUC	*k*	AUC	*k*	AUC	*k*	AUC	*k*	AUC
1	0,8377344	21	0,8170313	41	0,8000000	61	0,8602848	81	0,8503165
2	0,8118750	22	0,8252344	42	0,8441456	62	0,7310127	82	0,8091772
3	0,7555469	23	0,8687500	43	0,8852057	63	0,8666139	83	0,8141614
4	0,8117969	24	0,8952344	44	0,8049051	64	0,7719937	84	0,8378956
5	0,8450000	25	0,7835938	45	0,8014241	65	0,9166139	85	0,8523734
6	0,8209375	26	0,8509375	46	0,7912975	66	0,8059335	86	0,8240506
7	0,8271094	27	0,8342969	47	0,8490506	67	0,8940665	87	0,8289557
8	0,7751563	28	0,8371094	48	0,8133703	68	0,8757120	88	0,8402690
9	0,8564844	29	0,7728125	49	0,7977057	69	0,8515032	89	0,7981804
10	0,8450000	30	0,8300000	50	0,7841772	70	0,8590190	90	0,8678797
11	0,8105469	31	0,8185938	51	0,7989715	71	0,8077532	91	0,8606013
12	0,8807031	32	0,8186719	52	0,8411392	72	0,8328323	92	0,7956487
13	0,8328125	33	0,8428125	53	0,7584652	73	0,7608386	93	0,7628956
14	0,8278906	34	0,8180469	54	0,8366297	74	0,8102848	94	0,8756329
15	0,8875000	35	0,8277344	55	0,7957278	75	0,7791139	95	0,8302215
16	0,8103125	36	0,8025000	56	0,8424051	76	0,8550633	96	0,8814873
17	0,8866406	37	0,8316406	57	0,8704905	77	0,8273734	97	0,8454114
18	0,8340625	38	0,8354688	58	0,8525316	78	0,8439082	98	0,8821994
19	0,8590625	39	0,8069531	59	0,8497627	79	0,8279272	99	0,7945411
20	0,8330469	40	0,8500000	60	0,8166139	80	0,8456487	100	0,8209652

Prueba de normalidad

a) Formulación de hipótesis estadística

Hipótesis nula:

H_0: La variable AUC del modelo de Árbol de decisión proviene de una distribución normal.

Hipótesis alterna:

H_1: La variable AUC del modelo de Árbol de decisión no proviene de una distribución normal.

b) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

c) Elección del estadístico de contraste

Prueba de Lilliefors.

```
from statsmodels.stats.diagnostic import lilliefors

#Seleccionamos los AUC del modelo de árbol de decisión
dt = df_auc['Model'] == "Decision Trees-CART"
df_dt = df_auc[dt]
df_dt

lilliefors_test_auc = lilliefors(df_dt.AUC)
lilliefors_test_auc
print("statistic: "+str('{:.5f}'.format(lilliefors_test_auc[0])))
print("pvalue: "+str('{:.5f}'.format(lilliefors_test_auc[1])))
```

```
statistic: 0.05260
pvalue: 0.71356
```

Figura 28. Resultados de prueba de normalidad del AUC del modelo de Árbol de decisión

d) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Como $p - valor = 0.71356 \geq 0.05$, no se rechaza H_0, por lo tanto, la variable de estudio AUC proviene de una distribución normal, para un nivel de confianza del 95%.

Contrastación de hipótesis

a) Hipótesis específica 2

El modelo construido con la técnica de árboles de decisión obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

b) Formulación de hipótesis estadística

Hipótesis nula:

H_0: $\mu \leq 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de árboles de decisión es menor o igual a 0.74.

Hipótesis alterna:

H_1: $\mu > 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de árboles de decisión es mayor a 0.74.

c) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

d) Elección del estadístico de contraste

Elegimos la prueba paramétrica de T-Student para una sola muestra, por cuanto el conjunto de datos sigue una distribución normal y se desconoce la desviación estándar poblacional. El cálculo del valor de este estadístico lo realizamos con la librería pcipy de Python. La Figura 29 muestra el valor de este estadístico.

```
from scipy import stats
stats.ttest_1samp(df_dt.AUC, popmean=0.75, alternative='greater')
```

```
Ttest_1sampResult(statistic=23.136422648441712, pvalue=5.112209477537844e-42)
```

Figura 29. Resultados de la prueba de hipótesis específica 2

e) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Dado que $p - valor$ =5.112209477537844e-42<0.05, entonces rechazamos H_0, y aceptamos H_1.

Existe evidencia estadística suficiente de que la media de las áreas bajo la curva ROC del modelo construido con la técnica de árboles de decisión es mayor a 0.74, con un nivel de confianza del 95%.

Por lo tanto, se comprueba la hipótesis específica 2, que indica: El modelo construido con la técnica de árboles de decisión obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

4.4. Predecir utilizando clasificadores combinados la satisfacción laboral de docentes de educación básica del Perú

Para el logro de este objetivo se emplearon tres algoritmos pertenecientes a los clasificadores combinados o algoritmos de ensamble, los cuales son el algoritmo de Bosques aleatorios, XGBoost y Gradient boosting.

La Figura 30 muestra el código fuente para la creación de los modelos y el ajuste con el conjunto de datos de entrenamiento.

```
from sklearn.ensemble import RandomForestClassifier
rf_cl=RandomForestClassifier(max_depth=6, n_estimators=100,random_state=42)
rf_cl.fit(X_sm,Y_sm)
score=rf_cl.score(X_sm,Y_sm)
print("Exactitud: "+str('{:.5f}'.format(score)))
```

```
Exactitud: 0.76951
```

(a)

```
import xgboost as xgb
xgb_cl=xgb.XGBClassifier(max_depth=6, n_estimators=100,random_state=42,verbosity=0)
xgb_cl.fit(X_sm,Y_sm)
score=xgb_cl.score(X_sm,Y_sm)
print("Exactitud: "+str('{:.5f}'.format(score)))
```

```
Exactitud: 0.85402
```

(b)

```
from sklearn.ensemble import GradientBoostingClassifier
gb_cl = GradientBoostingClassifier(n_estimators=100, max_depth=6, random_state=42)
gb_cl.fit(X_sm, Y_sm)
score=gb_cl.score(X_sm,Y_sm)
print("Exactitud: "+str('{:.5f}'.format(score)))
```

```
Exactitud: 0.83557
```

(C)

Figura 30. Script para la construcción de los modelos de clasificadores combinados

La Figura 31, muestra las matrices de confusión obtenidas con los modelos de Bosques aleatorios, XGBoost y *Grandient Boosting* para los conjuntos de datos de entrenamiento y prueba. Los ítems a y b corresponden a la matriz de confusión del modelo de Bosques aleatorios en ambos conjuntos de datos, donde se observa que el ítem b contiene los valores más bajos en falsos negativos (docentes que fueron clasificados como satisfechos cuando en realidad estos se encuentran insatisfechos) 163(4.08%), este valor se representa en dicha matriz como FN; además, se evidencia los valores más altos en verdaderos positivos (docentes que fueron clasificados como

insatisfechos cuando en realidad eran docentes insatisfechos) 484 (12.13%) que se representan en dicha matriz como VP. Estos valores de la matriz de confusión pueden ser tomados como indicadores de que el modelo tiene un buen acierto en la clasificación e identificación de la clase positiva que para el presente estudio esta clase es 1 y representa a docentes insatisfechos (Fallucchi *et al.*, 2020).

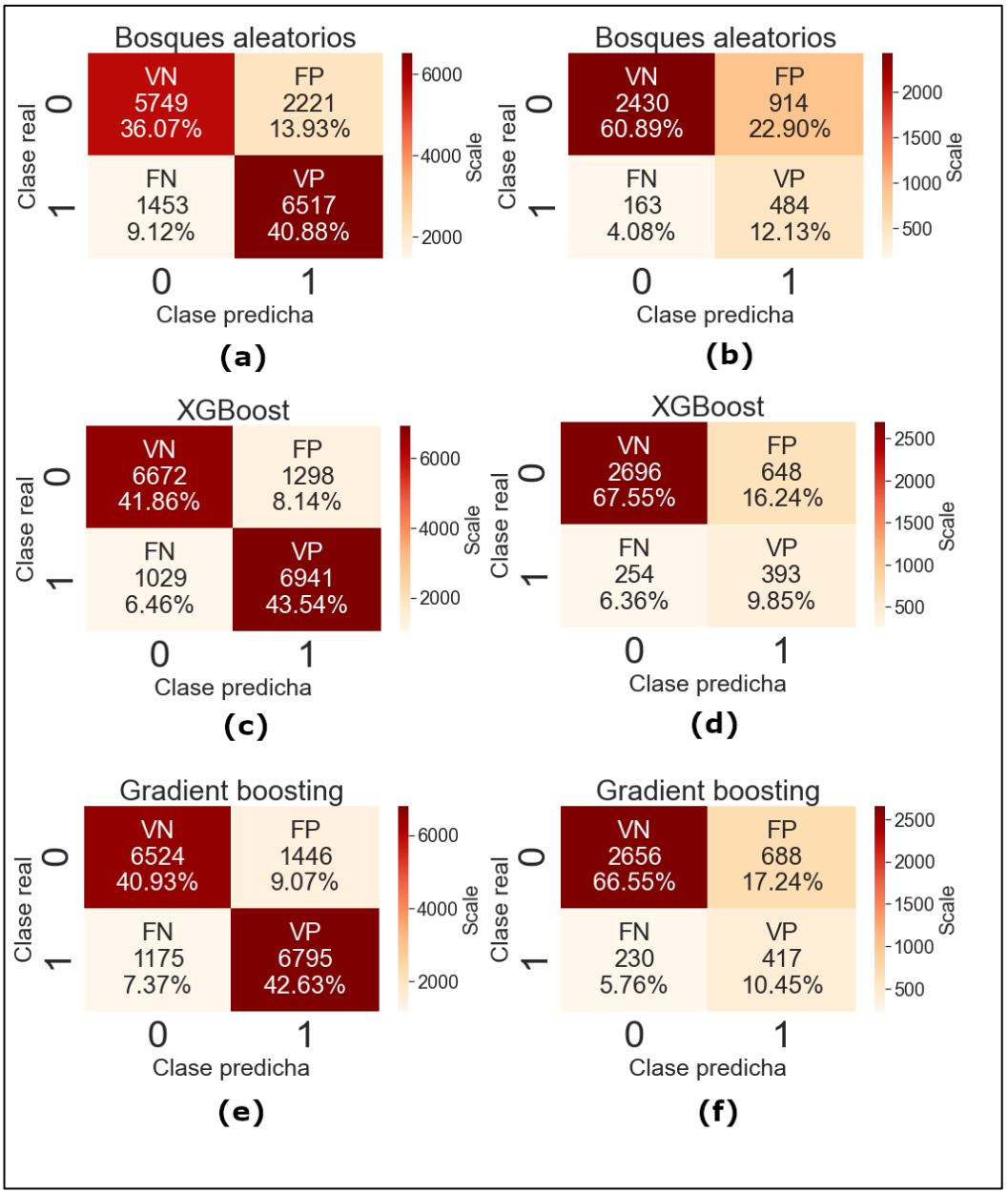


Figura 31. Matriz de confusión de los modelos de clasificadores combinados

Con los valores de la matriz de confusión de la Figura 31, se procede a calcular las métricas para los tres modelos, estas métricas se mostrarán en la Tabla 8.

Tabla 8

Métricas para los modelos de clasificadores combinados

Modelo	Métrica	Conjunto de datos de entrenamiento	Conjunto de datos de prueba
Bosques aleatorios	Exactitud	0,76775	0,73215
	Sensibilidad	0,81393	**0,74807**
	Especificidad	0,72158	0,72907
	Valor predictivo positivo	0,74512	0,34820
	Valor predictivo negativo	0,79500	**0,93733**
	F1 Score	0,77800	0,47521
	AUC	0,85206	**0,82001**
XGBoost	Exactitud	0,85402	**0,77399**
	Sensibilidad	0,87089	0,60742
	Especificidad	0,83714	**0,80622**
	Valor predictivo positivo	0,84246	**0,37752**
	Valor predictivo negativo	0,86638	0,91390
	F1 Score	0,85644	0,46564
	AUC	0,93579	0,79743
Gradient boosting	Exactitud	0,83557	0,76998
	Sensibilidad	0,85257	0,64451
	Especificidad	0,81857	0,79426
	Valor predictivo positivo	0,82454	0,37738
	Valor predictivo negativo	0,84738	0,92030
	F1 Score	0,83832	**0,47603**
	AUC	0,91754	0,80759

La Tabla 8 muestra las métricas de desempeño de los tres modelos de clasificadores combinados en los conjuntos de datos de entrenamiento y prueba, estos resultados evidencian que el modelo XGBoost obtiene el mayor valor de exactitud con un 77.40 %, mejor valor de especificidad del 80.62 % y valor predictivo positivo más alto del 0.377. En relación a la sensibilidad, valor predictivo negativo y área bajo la curva ROC (AUC), el algoritmo de Bosques Aleatorios obtiene los mejores valores del 74.81%, 93.73% y 0.8200 respectivamente. El valor más alto de F1-Score es de 0.4760 y fue obtenido por el algoritmo *Gradient Boosting*. Siendo la sensibilidad el porcentaje de instancias clasificadas correctamente de la clase positiva o clase de interés (docente insatisfecho) y además encontrándose un desbalanceo parcial en el conjunto de datos de prueba, donde la métrica de exactitud no es del todo confiable. Se determina que el modelo de Bosques aleatorios tiene mejor desempeño para los clasificadores combinados en las métricas de sensibilidad y AUC, esto indica que el mencionado modelo identifica mejor a docentes insatisfechos, además un AUC del 0.82 (Figura 32

ítem b) indica que la capacidad para discriminar la clase insatisfecho de la clase satisfecho es buena (Gironés *et al.*, 2017). Otra métrica importante a considerar cuando la variable objetivo se encuentra desbalanceada es la puntuación F1 o F1-Score. Este valor fue muy similar en los modelos de Gradient Boosting y Bosques aleatorios.

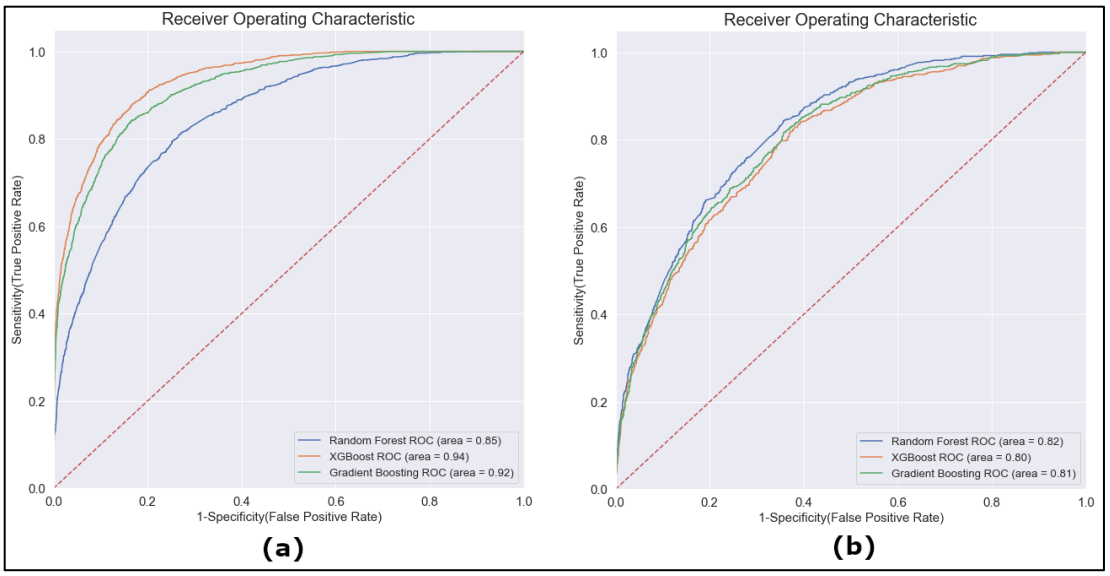


Figura 32. Área bajo la curva ROC (AUC) de los modelos de clasificadores combinados. (a) Gráfica del área bajo la curva (ROC) en el conjunto de datos de entrenamiento. (b) Gráfica del área bajo la curva (ROC) en el conjunto de datos de prueba.

Análisis inferencial

Se procede a realizar el análisis inferencial con los valores AUC obtenidos con el modelo de Bosques aleatorios que se muestran en la Tabla 9.

Tabla 9

Valores AUC del modelo de Bosques aleatorios

k	AUC	*k*	AUC	*k*	AUC	*k*	AUC	*k*	AUC
1	0,8214844	21	0,8447656	41	0,8136867	61	0,8682753	81	0,8636867
2	0,8430469	22	0,8474219	42	0,8525316	62	0,7417722	82	0,8540348
3	0,7976563	23	0,8778906	43	0,9048259	63	0,8704905	83	0,8431171
4	0,8197656	24	0,9193750	44	0,8772943	64	0,8060127	84	0,8464399
5	0,8436719	25	0,8303906	45	0,8344937	65	0,9155063	85	0,8911392
6	0,8385156	26	0,8508594	46	0,8363924	66	0,8303006	86	0,8526899
7	0,8421875	27	0,8467188	47	0,8686709	67	0,9371044	87	0,8444620
8	0,8097656	28	0,8528906	48	0,8218354	68	0,8472310	88	0,8443038
9	0,8719531	29	0,7946875	49	0,8300633	69	0,8605222	89	0,8067247
10	0,8525781	30	0,8407813	50	0,8084652	70	0,8897943	90	0,8810918
11	0,8489063	31	0,8455469	51	0,8113133	71	0,8120253	91	0,8924842
12	0,8975781	32	0,8532813	52	0,8346519	72	0,8338608	92	0,8332278
13	0,8384375	33	0,8642188	53	0,7742880	73	0,7623418	93	0,7987342
14	0,8723438	34	0,8152344	54	0,8466772	74	0,8257911	94	0,8984177
15	0,9068750	35	0,8339063	55	0,8268987	75	0,8097310	95	0,8416139
16	0,8230469	36	0,8263281	56	0,8298259	76	0,8518196	96	0,8734968
17	0,8821094	37	0,8225000	57	0,8961234	77	0,8394778	97	0,8567247
18	0,8518750	38	0,8600781	58	0,8641614	78	0,8556171	98	0,8874209
19	0,8756250	39	0,8310938	59	0,8678006	79	0,8524525	99	0,8004747
20	0,8214844	40	0,8512500	60	0,8204905	80	0,8428006	100	0,8085443

Prueba de normalidad

a) Formulación de hipótesis estadística

Hipótesis nula:

H_0: La variable AUC del modelo de Bosques aleatorios proviene de una distribución normal.

Hipótesis alterna:

H_1: La variable AUC del modelo de Bosques aleatorios no proviene de una distribución normal.

b) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

c) Elección del estadístico de contraste

Prueba de Lilliefors.

```
from statsmodels.stats.diagnostic import lilliefors

#Seleccionamos los AUC del modelo de Random forest
dt = df_auc['Model'] == "Random Forest"
df_rf = df_auc[dt]
df_rf

lilliefors_test_auc = lilliefors(df_rf.AUC)
lilliefors_test_auc
print("statistic: "+str('{:.5f}'.format(lilliefors_test_auc[0])))
print("pvalue: "+str('{:.5f}'.format(lilliefors_test_auc[1])))
```

```
statistic: 0.08832
pvalue: 0.05410
```

Figura 33. Resultados de prueba de normalidad del AUC del modelo de Bosques aleatorios

d) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Como $p - valor = 0.05410 \geq 0.05$, no se rechaza H_0, por lo tanto, la variable de estudio AUC proviene de una distribución normal, para un nivel de confianza del 95%.

Contrastación de hipótesis

a) Hipótesis específica 3

El modelo construido con la técnica de clasificadores combinados Random Forest obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

b) Formulación de hipótesis estadística

Hipótesis nula:

H_0: $\mu \leq 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de clasificadores combinados Random Forest es menor o igual a 0.74.

Hipótesis alterna:

H_1: $\mu > 0.74$ La media de las áreas bajo la curva ROC del modelo construido con la técnica de clasificadores combinados Random Forest es mayor a 0.74.

c) Nivel de significancia

Establecemos el nivel de significancia $\alpha = 5\% = 0.05$

d) Elección del estadístico de contraste

Elegimos la prueba paramétrica de T-Student para una sola muestra, por cuanto el conjunto de datos sigue una distribución normal y se desconoce la desviación estándar poblacional. El cálculo del valor de este estadístico lo realizamos con la librería Scipy de Python. La Figura 34 muestra el valor de este estadístico.

```
from scipy import stats
stats.ttest_1samp(df_rf.AUC, popmean=0.75, alternative='greater')
```

```
Ttest_1sampResult(statistic=29.221760767574995, pvalue=8.834113520893689e-51)
```

Figura 34. Resultados de la prueba de hipótesis específica 3

e) Establecimiento de la regla de decisión

Rechazar H_0, si $p - valor < \alpha$

No rechazar H_0, si $p - valor \geq \alpha$

Dado que $p - valor$ =8.834113520893689e-51<0.05, entonces rechazamos H_0, y aceptamos H_1.

Existe evidencia estadística suficiente de que la media de las áreas bajo la curva ROC del modelo construido con la técnica de clasificadores combinados Random forest es mayor a 0.74, con un nivel de confianza del 95%.

Por lo tanto, se comprueba la hipótesis específica 3, que indica: El modelo construido con la técnica de clasificadores combinados Random Forest obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

4.5. Comparar las métricas de la técnica de regresión logística respeto a la técnica de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

Para el cumplimiento de este objetivo se evaluó el desempeño de los cinco algoritmos de aprendizaje automático en la predicción de la satisfacción laboral del docente de educación básica. Estos algoritmos son: Regresión logística, *Gradient boosting*, *Random forest*, *XGBoost* y *Decision trees-CART*. En la Tabla 10, se observa los resultados obtenidos tras aplicar la técnica K-Fold Cross-Validation con k=100 sobre el conjunto de datos de entrenamiento. Obsérvese que el modelo XGBoost obtuvo mejor valor en todas las métricas, una exactitud con una media de 0.828 ± 0.031, una sensibilidad media de 0.857 ± 0.042, un valor predictivo positivo medio de 0.810 ± 0.034, F1-Score del 0.832 ± 0.030 y un AUC 0.906 ± 0.023.

Tabla 10

Resumen estadístico de métricas en el conjunto de datos de entrenamiento

Modelo	Exactitud	Sensibilidad	VPP	F1-Score	AUC
Árboles de Decisión-CART	0,748±0.034	0,698±0.052	0,776±0.037	0,734±0.039	0,829±0.034
Gradient Boosting	0,810±0.033	0,840±0.044	0,794±0.036	0,816±0.033	0,892±0.026
Regresión Logística	0,736±0.033	0,748±0.053	0,732±0.034	0,739±0.035	0,819±0.031
Random Forest	0,762±0.034	0,812±0.051	0,739±0.035	0,773±0.033	0,846±0.033
XGBoost	0,828±0.031	0,857±0.042	0,810±0.034	0,832±0.030	0,906±0.023

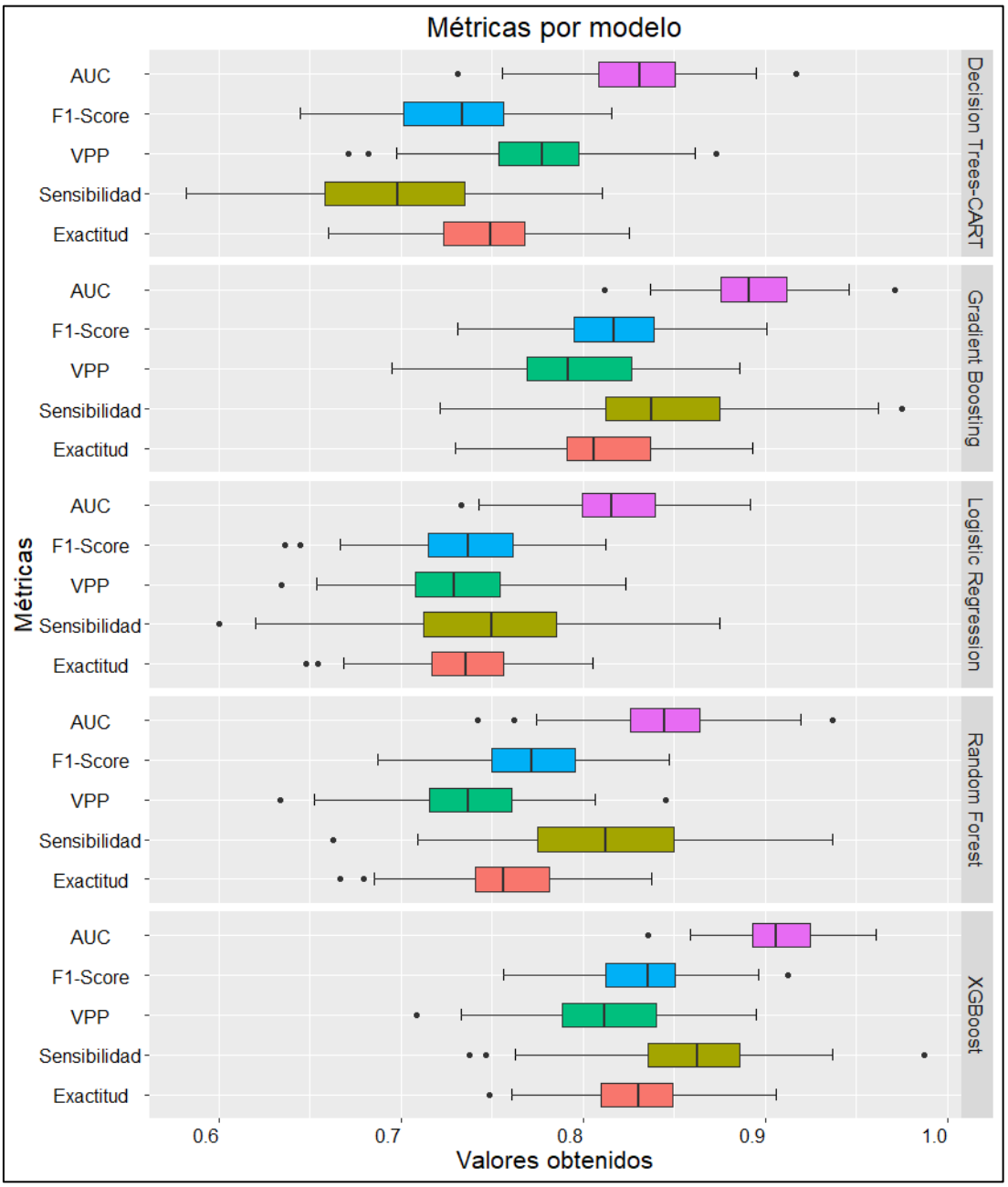


Figura 35. Métricas en el conjunto de datos de entrenamiento

En la Figura 35, se observa el diagrama de cajas y bigotes de las métricas obtenidas en el conjunto de datos de entrenamiento. Analizando la mediana podemos afirmar que el algoritmo con peores valores de exactitud, valor predictivo positivo o VPP y AUC es el de Regresión logística, en relación a la métrica de sensibilidad el algoritmo con los valores más bajos es Árboles de decisión-CART.

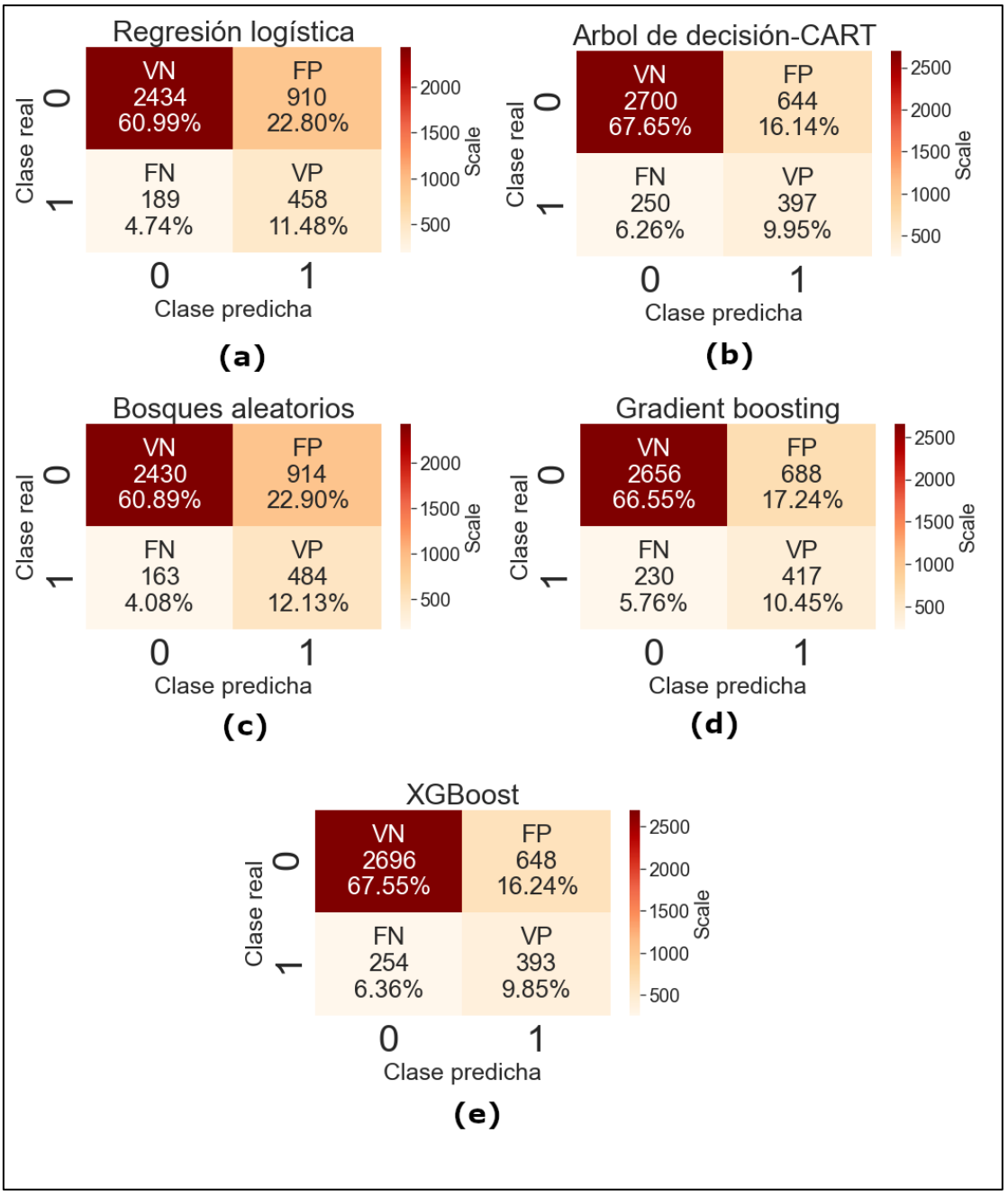


Figura 36. Matriz de confusión de modelos estudiados en el conjunto de datos de prueba

La Figura 36, muestra la matriz de confusión de los modelos en estudio en el conjunto de datos de prueba. En la matriz de confusión del modelo de Bosques aleatorios o

Random Forest (Figura 36 ítem c) se observa los valores más bajos de falsos negativos 163 (docentes insatisfechos, pero no están clasificados como tales) y los valores más altos de verdaderos positivos 484 (docentes insatisfechos que el modelo clasificó como tales).

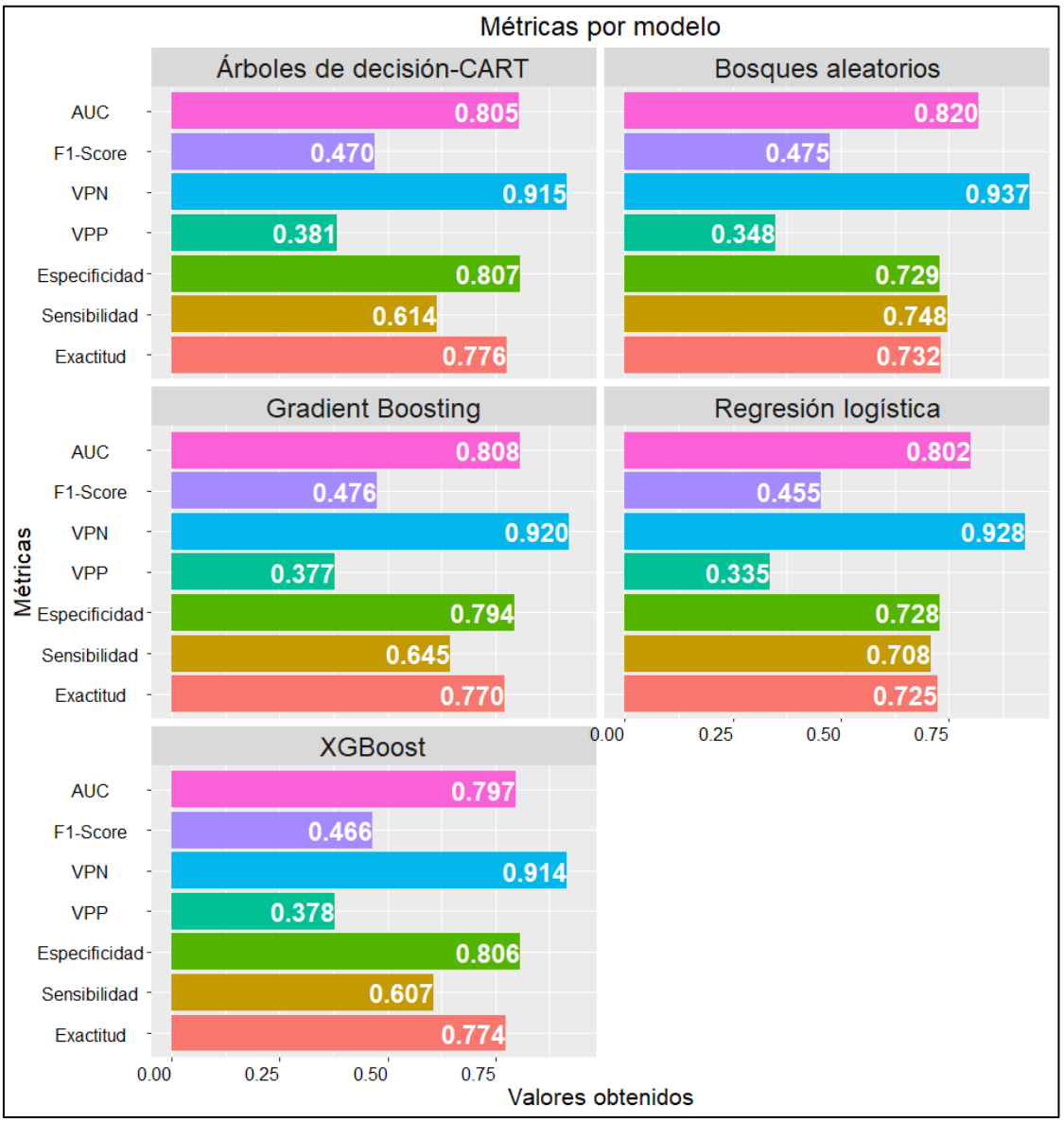


Figura 37. Métricas obtenidas en el conjunto de datos de prueba

La Figura 37, muestra las métricas que se obtuvieron luego de la predicción del conjunto de datos de prueba. Estás métricas serán consideradas para la selección del mejor modelo para la predicción de la satisfacción laboral docente de educación básica. Obsérvese que el valor de la exactitud más alta se obtuvo con el algoritmo Árboles de decisión-CART del 77.6%.

Dado que el conjunto de datos se encuentra parcialmente desbalanceada este indicador podría ser engañoso. En este escenario es importante una métrica que nos permita analizar la predicción de la clase de interés (1: Insatisfecho), por ello optamos por la sensibilidad. El modelo construido con el algoritmo de Bosques aleatorios o Random Forest obtuvo el valor de sensibilidad de 74.8%, siendo el más alto en comparación a los demás.

En la Figura 37 también se observa que los algoritmos de Árboles de decisión-CART y XGBoost obtuvieron valores de especificidad más altos del 80.7% y 80.6% respectivamente, indicando así que estos son buenos para predecir la clase 0, es decir docentes satisfechos. Otra métrica en este contexto, que permite conocer la probabilidad de que el docente esté insatisfecho dado que fue clasificado como tal, es el valor predicativo positivo, este valor fue mayor en el modelo construido por el algoritmo Árboles de decisión-CART con un 38.1%.

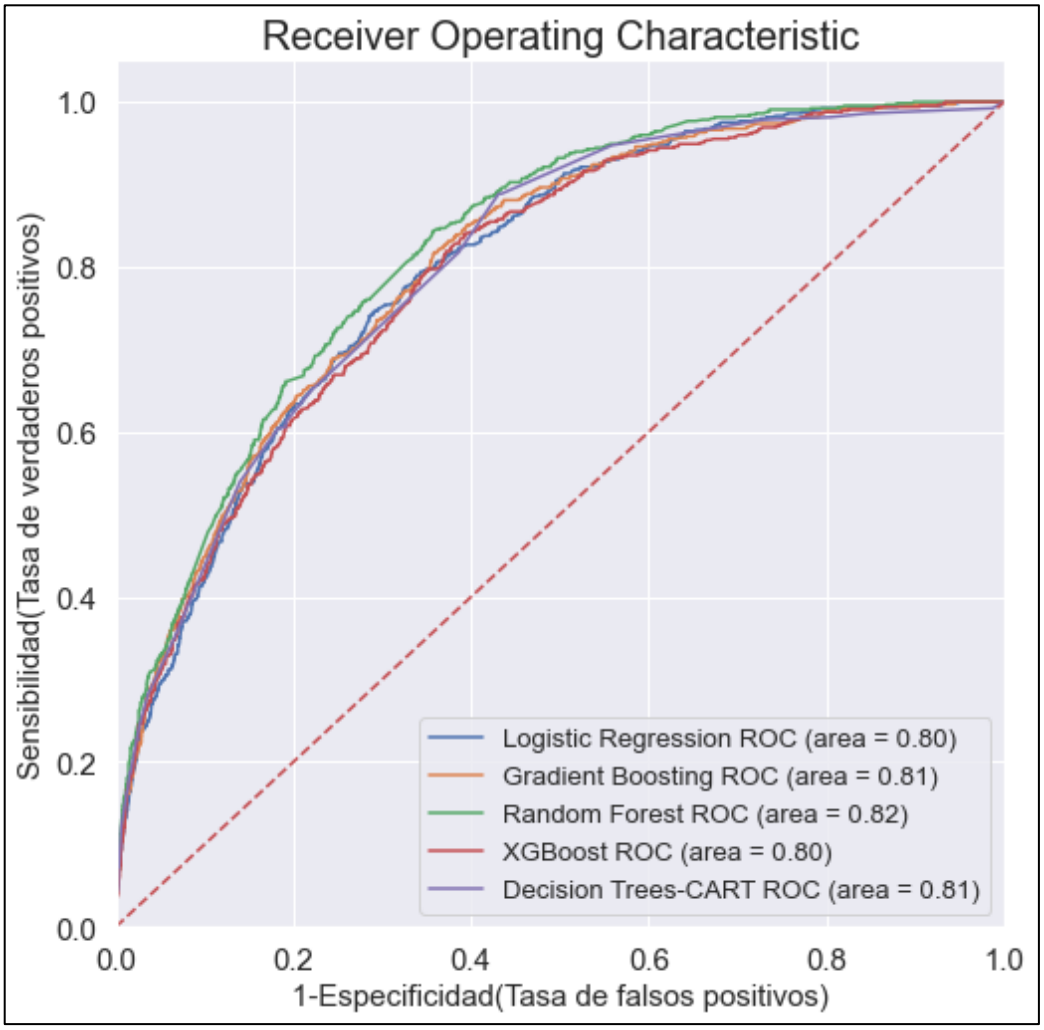


Figura 38. Comparativa de área bajo la curva ROC (AUC)

La Figura 38, muestra el área bajo la curva ROC obtenido luego de la predicción en el conjunto de datos de prueba, donde el modelo construido con el algoritmo Random Forest obtiene un valor de AUC=0.82, seguido por los algoritmos de Gradient Boosting y árboles de decisión con valores de AUC=0.81, demostrando de esta manera que la capacidad discriminativa el modelo de Random Forest para la clase insatisfecho y satisfecho es buena, se afirma ello de acuerdo a la escala de valoración propuesta por (Gironés et al., 2017).

4.6. Discusión de los resultados

Tras los hallazgos encontrados en la presente investigación, aceptamos la hipótesis que establece que los modelos predictivos *machine learning* de regresión logística, árboles de decisión y clasificadores combinados obtienen una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

En relación al valor de la métrica de exactitud del 77.6%, obtenido por el algoritmo de Árboles de decisión-CART en el conjunto de datos de prueba. Se percibe que estos hallazgos son similares a los resultados reportados por (Tao Chen et al., 2021) donde el mencionado algoritmo obtiene una exactitud del 76.15% en la predicción de la satisfacción laboral. Los investigadores emplearon un conjunto de datos de 280 instancias con 20 variables independientes. La técnica de selección de características para variables categóricas que emplearon los mencionados investigadores es la misma que se utilizó en la presente investigación.

Respecto a la métrica de sensibilidad del 74.8% obtenida por el algoritmo de Bosques aleatorios o Random Forest, de acuerdo con la revisión del estado del arte, se observa que este algoritmo obtiene los mejores valores en esta métrica, por ejemplo los resultados de Rustam *et al.* (2021) demuestran que el algoritmo Random Forest obtiene el valor más alto de sensibilidad del 85% en comparación de los algoritmos de Regresión logística, Máquinas de soporte vectorial, Gradient boosting, Extreme gradient boosting, Perceptrón multicapa. En el estudio de referencia, Rustam *et al.* (2021) emplearon el conjunto de datos provenientes de reseñas de texto de los empleados de Google, Facebook, Amazon, Microsoft y Apple y la selección de características se realizó con la técnica denominada "Term Frequency — Inverse Document Frequency". Otro estudio que reporta similares resultados en esta métrica es el de Arambepola y Munasinghe (2021), donde este algoritmo obtiene el valor más

alto de sensibilidad de 80% comparado con los algoritmos de Máquina de soporte vectorial, Regresión logística y Redes neuronales que obtienen un 79% en la predicción de la satisfacción laboral de los profesionales de tecnologías de información.

CONCLUSIONES

Tras la obtención de resultados, el análisis e interpretación de mismos, a la luz de las evidencias reportadas en el presente estudio, se concluye que:

- El presente estudio permitió analizar e identificar los predictores más importantes y confiables de la satisfacción laboral de docentes de educación básica empleando datos procedentes de la encuesta nacional docente ENDO-2018 elaborada por el Ministerio de Educación del Perú. Debido a que el conjunto de datos original poseía variables de tipo numérico y categóricos fue necesario tratamientos diferenciados. En razón a esto, se realizó la selección de características empleando el filtro ANOVA F-test para las variables numéricas y el filtro Chi-Cuadrado para las variables categóricas. Los resultados evidencian que las variables más importantes en la predicción de la satisfacción laboral de docentes son: ingresos económicos, la satisfacción con la vida, con la autoestima, con la actividad pedagógica, con la relación con el director(a), percepción de las condiciones de vida, satisfacción con sus relaciones familiares, problema de salud relacionado con la depresión y la satisfacción de la relación con sus colegas.

- Se logró establecer el modelo de regresión logística para la satisfacción laboral de docentes de educación básica del Perú, $p(y=1|X) = \frac{e^{(6.47-0.29*P401\_12\_No-0.44*P501\_A-0.33*P509-0.56*P818\_1-0.13*P818\_6-0.36*P819\_1-0.18*P819\_4-0.71*P819\_5-1.29*P819\_8)}}{1+e^{(6.47-0.29*P401\_12\_No-0.44*P501\_A-0.33*P509-0.56*P818\_1-0.13*P818\_6-0.36*P819\_1-0.18*P819\_4-0.71*P819\_5-1.29*P819\_8)}}$ el mismo que resultó ser significativo con un *Likelihood ratio p-value=0.000*. Para este modelo, las variables que resultaron ser significativas en la predicción de la satisfacción laboral de docentes de educación básica fueron: ingresos totales, percepción con las condiciones de vida, satisfacción con la vida, satisfacción con su autoestima, satisfacción con su actividad pedagógica, satisfacción de la relación con sus colegas, satisfacción de la relación con el/la director(a), la satisfacción con su salario y problemas relacionados con la depresión. Los resultados de las mediciones demuestran que este modelo logra una exactitud del 72.5 %, sensibilidad del 70.8 % y un área bajo la curva ROC del 0.80, esta área bajo la curva ROC es superior al 0.74; por lo tanto, se concluye que el modelo construido con la técnica de regresión logística tiene una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

- Se logró predecir utilizando árboles de decisión la satisfacción laboral de docentes de educación básica del Perú. Los resultados de las mediciones demuestran que este modelo logra una exactitud del 77.6 %, sensibilidad del 61.4 % y un área bajo la curva ROC del 0.81, esta área bajo la curva ROC es superior al 0.74; por lo tanto, se concluye que el modelo construido con la técnica de Árboles de decisión tiene una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

- Se logró predecir utilizando clasificadores combinados la satisfacción laboral de docentes de educación básica del Perú, para esto se emplearon los algoritmos de Bosques aleatorios, Gradient boosting y XGBoot. Los resultados de las mediciones demuestran que el modelo construido con el algoritmo de Bosques aleatorios obtiene una exactitud del 73.2 %, una sensibilidad del 74.8 % y un área bajo la curva del 0.82, esta área bajo la curva ROC es superior al 0.74; por lo tanto, se concluye que el modelo construido con la técnica de clasificadores combinados Random Forest tiene una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.

- Tras la evaluación y comparación de los modelos construidos por los algoritmos de Regresión logística, Gradient Boosting, Random Forest, XGBoost y Árboles de decisión-CART, el algoritmo que obtuvo los mejores resultados en el conjunto de datos analizados fue Random Forest, el mismo que obtuvo sensibilidad del 74.8 %. Esta métrica representa la capacidad del modelo para clasificar correctamente a una instancia de la clase positiva (1: insatisfecho), un área bajo la curva ROC de 0.82, menor valor de falsos negativos 163 y mayor valor de verdaderos positivos 484 en la matriz de confusión. Debido a que la distribución de las clases (satisfecho e insatisfecho) en la variable objetivo del conjunto de datos de prueba se encuentra desequilibrada, no se consideró la métrica de exactitud o accuracy como determinante para la selección del mejor algoritmo.

RECOMENDACIONES

Concluida la presente investigación, y de acuerdo con las limitaciones y restricciones que se presentaron, proponemos las siguientes recomendaciones:

- En la presente investigación se identificó las variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica, empleando métodos de filtro como ANOVA-F test y Chi-cuadrado para la selección de características. Se recomienda el uso de métodos de envoltura como la eliminación recursiva de características (RFE), que no fue posible aplicar por la limitante de recurso computacional.

- Ante la gran dimensionalidad del conjunto de datos, resulta una buena práctica realizar el test de significancia estadística de los predictores antes de la construcción del modelo de Regresión logística debido a que permite optimizar la selección de variables.

- Se recomienda el empleo de alguna técnica de ajuste de hiperparámetros como búsqueda en cuadricula o búsqueda aleatoria para encontrar el valor más adecuado de la profundidad máxima o *max_depth* del modelo de árbol de decisión, debido a que valores inadecuados de este, pueden ocasionar un sobreajuste del modelo. En el presente estudio se tuvo limitaciones de recurso computacional para la aplicación de las mencionadas técnicas.

- Para una mejor predicción de la satisfacción laboral de docentes de educación básica mediante clasificadores combinados, se recomienda el uso tecnologías de *cloud computing* que provean de hardware y software suficientes para las necesidades de dichos algoritmos.

- Se espera que los resultados del presente estudio puedan servir como punto de partida para la elaboración de modelos predictivos de clasificación de la satisfacción laboral de docentes cada vez más óptimos, empleando otros algoritmos más sofisticados como las redes neuronales artificiales.

BIBLIOGRAFÍA

Alshawabkeh, M., Jang, B., & Kaeli, D. (2010). Accelerating the local outlier factor algorithm on a GPU for intrusion detection systems. *International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS*, 104–110. https://doi.org/10.1145/1735688.1735707

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Questiio*, *25*(3), 479–498.

Amat-Rodrigo, J. (2016). *Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping*.

Angulo-Paredes, S. A., Fuster-Guillén, D., Castro, A. S., Rodríguez, E. L. B., & Ramírez, T. V. C. (2021). Características predominantes del aprendizaje organizacional que influyen en el bienestar laboral de los docentes del Perú. *Propósitos y Representaciones*, *9*(1), 1035. https://doi.org/http://dx.doi.org/10.20511/pyr2021

Aouadni, I., & Rebai, A. (2016). Decision support system based on genetic algorithm and multi-criteria satisfaction analysis (MUSA) method for measuring job satisfaction. *Annals of Operations Research 2016 256:1*, *256*(1), 3–20. https://doi.org/10.1007/S10479-016-2154-Z

Arambepola, N., & Munasinghe, L. (2021a). What makes job satisfaction in the information technology industry? *2021 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 99–105. https://doi.org/10.1109/SCSE53661.2021.9568331

Arambepola, N., & Munasinghe, L. (2021b). What makes job satisfaction in the information technology industry? *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2021*, 99–105. https://doi.org/10.1109/SCSE53661.2021.9568331

Araoz-Estrada, E. G., & Ramos-Gallegos, N. A. (2021). Satisfacción laboral y compromiso organizacional en docentes de la amazonía peruana. *Educação Formação*, *6*(1), e3854. https://doi.org/10.25053/redufor.v6i1.3854

Arias, F. (2016). *El Proyecto de lnvestigación lntroducción a la metodologío científico* (Episteme (Ed.); 7ma Edició). https://kupdf.net/queue/el-proyecto-de-investigacion-fidias-arias-7ma-edic-2016pdf_5a1b4afde2b6f5e526da642c_pdf?queue_id=-1&x=1646192537&z=MjgwMDpiZjA6MjQwNDoxMjQ4OmM0YmE6Y2UyYzpmZDdiOjU3NGE=

Asadujjaman, M. D., Rashid, M. H. O., Nayon, M. A. A., Biswas, T. K., Arani, M., & Billal,

M. M. (2020). Teachers' job satisfaction at tertiary education: A case of an engineering university in Bangladesh. *Proceedings of the International Conference on E-Learning, ICEL, 2020-Decem*, 238–242. https://doi.org/10.1109/ECONF51404.2020.9385512

Beunza-Nuin, J. J., Puertas-Sanz, E., & Condés-Moreno, E. (2020). *Inteligencia Artificial en entornos sanitarios. Tipos de algoritmos de "machine learning"* (Elseviers (Ed.)). Elsevier Connect.

Bourel, M., Segura, A. M., Crisci, C., López, G., Sampognaro, L., Vidal, V., Kruk, C., Piccini, C., & Perera, G. (2021). Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Research*, *202*, 117450. https://doi.org/10.1016/J.WATRES.2021.117450

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. (1987). Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, $27.95. *Cytometry*, *8*(5), 534–535. https://doi.org/10.1002/CYTO.990080516

Brownlee, J., Sanderson, M., Koshy, A., Cheremskoy, A., & Halfyard, J. (2020). *Machine Learning Mastery With Python: Data Cleaning, Feature Selection, and Data Transforms in Python*.

Bunge, M. (1999). BUSCAR LA FILOSOFÍA EN LAS CIENCIAS SOCIALES. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9). https://doi.org/10.1017/CBO9781107415324.004

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. In *Cell* (Vol. 173, Issue 7, pp. 1581–1592). Cell Press. https://doi.org/10.1016/j.cell.2018.05.015

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista Chilena de Infectología*, *29*(2), 138–141. https://doi.org/10.4067/S0716-10182012000200003

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/J.COMPELECENG.2013.11.024

Chen, Tao, Cao, Z., & Cao, Y. (2021). Comparison of Job Satisfaction Prediction Models for Construction Workers: CART vs. Neural Network. *Tehnički Vjesnik*, *28*(4), 1174–1181. https://doi.org/10.17559/TV-20200507022306

Chen, Tianqi, & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672

Chiang-Vega, M. M., Riquelme-Neira, G. R., & Rivas-Escobar, P. A. (2018). Relación entre Satisfacción Laboral, Estrés Laboral y sus Resultados en Trabajadores de una Institución de Beneficencia de la Provincia de Concepción. *Ciencia & Trabajo*, *20*(63), 178–186. https://doi.org/10.4067/s0718-24492018000300178

Cuesta, M., Fonseca-Pedrero, E., Vallejo, G., & Muñiz, J. (2013). Datos perdidos y propiedades psicométricas en los tests de personalidad. *Anales de Psicología*, *29*(1), 285–292. https://doi.org/10.6018/ANALESPS.29.1.137901

Dashdondov, K., Lee, S. M., & Kim, M. H. (2021). OrdinalEncoder and PCA based NB Classification for Leaked Natural Gas Prediction Using IoT based Remote Monitoring System. *Smart Innovation, Systems and Technologies*, *212*, 252–259. https://doi.org/10.1007/978-981-33-6757-9_32

Espinosa-Zúñiga, J. J., & Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, Investigación y Tecnología*, *21*(3), 1–16. https://doi.org/10.22201/FI.25940732E.2020.21.3.022

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 1–17. https://doi.org/10.3390/computers9040086

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*.

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, *13*(2–3), 99–115. https://doi.org/10.1016/s0167-739x(97)00015-0

Fisher, O. J., Watson, N. J., Escrig, J. E., Witt, R., Porcu, L., Bacon, D., Rigley, M., & Gomes, R. L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers and Chemical Engineering*, *140*, 106881. https://doi.org/10.1016/j.compchemeng.2020.106881

Gabrani, G., & Kwatra, A. (2018). Machine learning based predictive model for risk assessment of employee attrition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10963 LNCS*, 189–201. https://doi.org/10.1007/978-3-319-95171-3_16/COVER/

Gironés, J., Casas, J., Minguillón, J., & Caihuelas, R. (2017). *Minería de datos: modelos y algoritmos* (UOC (Ed.); Primera).

Greedy Function Approximation: A Gradient Boosting Machine on JSTOR. (n.d.). Retrieved January 30, 2022, from https://www.jstor.org/stable/2699986

Hair, J. F., Anderson, R. E., Tatham, R. L., Research, B. M., Black, W. C., Prentice, E., Cano, D., & Gómez Suárez, M. (1999). *ANÁLISIS MULTIVARIANTE* (P. HALL (Ed.); Quinta edi).

Han, J., Kamber, M., & Pei, J. (2012). *DATA MINING-Concepts and Techniques* (Elsevier (Ed.); Third Edit). Morgan Kaufmann.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 1–41. https://doi.org/10.1186/S40537-020-00305-W/FIGURES/4

Hassan, O., & Ibourk, A. (2021). Burnout, self-efficacy and job satisfaction among primary school teachers in Morocco. *Social Sciences & Humanities Open*, *4*(1), 100148. https://doi.org/10.1016/j.ssaho.2021.100148

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). Metodología de la investigación: las tres rutas cuantitativa, cualitativa y mixta. In *Mc Graw Hill* (Vol. 1, Issue Mexico).

Homocianu, D., Plopeanu, A. P., Florea, N., & Andries, A. M. (2020). Exploring the patterns of job satisfaction for individuals aged 50 and over from three historical regions of Romania. An inductive approach with respect to triangulation, cross-validation and support for replication of results. *Applied Sciences (Switzerland)*, *10*(7). https://doi.org/10.3390/APP10072573

Hong-Hua, M., Mi, W., Hong-Yun, L., & Yong-Mei, H. (2016). Influential factors of China's elementary school teachers' job satisfaction. *Springer Proceedings in Mathematics and Statistics*, *167*, 339–361. https://doi.org/10.1007/978-3-319-38759-8_26

Hossen, M. A., Hossain, E., Ishwar, A. K. Z., & Siddika, F. (2021). Ensemble method based architecture using random forest importance to predict employee's turn over. *Journal of Physics: Conference Series*, *1755*(1). https://doi.org/10.1088/1742-6596/1755/1/012039

Igual, L., & Seguí, S. (2020). *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer. https://doi.org/10.4018/978-1-7998-3053-5.ch001

Jain, R., & Nayyar, A. (2018). Predicting employee attrition using xgboost machine learning

approach. In IEEE (Ed.), *Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018* (pp. 113–120). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/SYSMART.2018.8746940

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1060–1073. https://doi.org/10.1016/J.JKSUCI.2019.06.012

Khera, S. N., & Divya. (2019). Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. *Vision*, *23*(1), 12–21. https://doi.org/10.1177/0972262918821221

Kuzey, C. (2018). Impact of Health Care Employees' Job Satisfaction on Organizational Performance Support Vector Machine Approach. *Journal of Economics and Financial Analysis*, *2*(1), 45–68. https://doi.org/10.1991/jefa.v2i1.a12

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Lee, A. N., & Nie, Y. (2014). Understanding teacher empowerment: Teachers' perceptions of principal's and immediate supervisor's empowering behaviours, psychological empowerment and work-related outcomes. *Teaching and Teacher Education*, *41*, 67–79. https://doi.org/10.1016/J.TATE.2014.03.006

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. *Frontiers in Genetics*, *10*, 1077. https://doi.org/10.3389/FGENE.2019.01077/BIBTEX

Liu, Y., Zhao, T., Ju, W., Shi, S., Shi, S., & Shi, S. (2017). Materials discovery and design using machine learning. In *Journal of Materiomics* (Vol. 3, Issue 3, pp. 159–177). Chinese Ceramic Society. https://doi.org/10.1016/j.jmat.2017.08.002

Lopes, J., & Oliveira, C. (2020). Teacher and school determinants of teacher job satisfaction: a multilevel analysis. *Https://Doi.Org/10.1080/09243453.2020.1764593*, *31*(4), 641–659. https://doi.org/10.1080/09243453.2020.1764593

Malander, N. M. (2016). Síndrome de Burnout y Satisfacción Laboral en Docentes de Nivel Secundario. *Ciencia & Trabajo*, *18*(57), 177–182. https://doi.org/10.4067/s0718-24492016000300177

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/J.PROCS.2016.07.111

MINEDU. (2022). *Ministerio de Educación del Perú | MINEDU*. http://www.minedu.gob.pe/politicas/docencia/encuesta-nacional-a-docentes-endo.php

Montgomery, D. (2004). *Diseño y análisis de experimentos* (G. N. Editores (Ed.); Segunda ed). Limusa wiley.

Moon, N. N., Mariam, A., Sharmin, S., Islam, M. M., Nur, F. N., & Debnath, N. (2021a). Machine learning approach to predict the depression in job sectors in Bangladesh. *Current Research in Behavioral Sciences*, *2*, 100058. https://doi.org/10.1016/j.crbeha.2021.100058

Moon, N. N., Mariam, A., Sharmin, S., Islam, M. M., Nur, F. N., & Debnath, N. (2021b). Machine Learning Approach to Predict the Depression in Job Sectors in Bangladesh. *Current Research in Behavioral Sciences*, 100058. https://doi.org/10.1016/J.CRBEHA.2021.100058

Moreno, M., Miguel, L., García, F., & Polo, J. (2001). Aplicación de técnicas de minería de datos en la construccion y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. *ADIS 2001*, *21812*, 1–16.

MUNEDU. (2018). *UE - ESCALE - Unidad de Estadística Educativa*.

Nader, M., Bernate, S. P. P., & Santa-Bárbara, E. S. (2014). Prediction of satisfaction and well-being at work: Towards a model of healthy organization in Colombia. *Estudios Gerenciales*, *30*(130), 31–39. https://doi.org/10.1016/j.estger.2014.02.006

Navarro-Pastor, J., & Losilla-Vidal, J. (2000). Psicothema - ANÁLISIS DE DATOS FALTANTES MEDIANTE REDES NEURONALES ARTIFICIALES. *Psicothema*, *12*(3), 503–510.

Padrón, J. (2016). *BASES DEL CONCEPTO DE INVESTIGAC*.

Panadero, E., & Alonso-Tapia, J. (2014). Teorías de autorregulación educativa: una comparación y reflexión teórica. *Psicologia Educativa*, *20*(1), 11–22. https://doi.org/10.1016/j.pse.2014.05.002

Park, Y., Seo, D. G., Park, J., Bettini, E., & Smith, J. (2016). Predictors of job satisfaction among individuals with disabilities: An analysis of South Korea's National Survey of employment for the disabled. *Research in Developmental Disabilities*, *53–54*, 198–212. https://doi.org/10.1016/j.ridd.2016.02.009

Parra, F. (2019). *Estadística y Machine Learning con R*.

Pazos, A., Pedreira, N., Rabuñal, J., & Pereira, J. (2007). *Inteligencia Artificial y Computación avanzada* (Fundación).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel,

M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez, C. (2008). *Técnicas de Análisis Multivariante de Datos* (Pearson (Ed.)).

Pérez, C., & Santín, D. (2008). *Minería de datos: técnicas y herramientas - César Pérez López - Google Libros* (T. E. Paraninfo (Ed.); 1ra ed.).

Ponce, J. C., Torres, A., Sprock, A. S., & Casali, A. (2014). *Inteligencia Artificial* (LATIN (Ed.); Primera Ed, Issue March). https://doi.org/10.13140/2.1.3720.0960

Potdar, K. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 975–8887.

Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing*, *9*(1), 49–66. https://doi.org/10.22364/BJMC.2021.9.1.04

Quintero, M. A., & Duran, M. (2004). Análisis del error tipo I en las pruebas de bondad de ajuste e independencia utilizando el muestreo con parcelas de tamaño variable (Bitterlich). *Bosque (Valdivia)*, *25*(3), 45–55. https://doi.org/10.4067/S0717-92002004000300005

Rainer, J., & Rodríguez, L. (2017). *Inteligencia Artificial Aplicada a la Defensa*.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (P. P. Ltd. (Ed.); Second edi).

Robbins, S. (2004). Comportamiento Organizacional, 10a. ed. In *Pearson Educación de México, S.A.*

Robertson, B. W., & Kee, K. F. (2017). Social media at work: The roles of job satisfaction, employment status, and Facebook use with co-workers. *Computers in Human Behavior*, *70*, 191–196. https://doi.org/10.1016/j.chb.2016.12.080

Rosado Gómez, A. A., & Verjel Ibáñez, A. (2015). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Tecnura*, *19*(45), 101–113. https://doi.org/10.14483/UDISTRITAL.JOUR.TECNURA.2015.3.A08

Rosati, G. (2017). Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning: Aplicación en la Encuesta permanente de hogares (EPH). *SaberEs*, *19*(1).

Ruiz-Quiles, M., Moreno-Murcia, J. A., & Vera-Lacárcel, J. A. (2015). Del soporte de autonomía y la motivación autodeterminada a la satisfacción docente. *European Journal of Education and Psychology*, *8*(2), 68–75.

https://doi.org/10.1016/j.ejeps.2015.09.002

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning Resumen Contexto Introducción. *XVIII Workshop de Investigadores En Ciencias de La Computación (Entre Ríos, Argentina)*, 131–134.

Rustam, F., Ashraf, I., Shafique, R., Mehmood, A., Ullah, S., & Sang Choi, G. (2021). Review prognosis system to predict employees job satisfaction using deep neural network. *Computational Intelligence*, *37*(2), 964–990. https://doi.org/10.1111/COIN.12440

Sadeghi, K., Ghaderi, F., & Abdollahpour, Z. (2021). Self-reported teaching effectiveness and job satisfaction among teachers: the role of subject matter and other demographic variables. *Heliyon*, *7*(6), e07193. https://doi.org/10.1016/j.heliyon.2021.e07193

Saisanthiya, D., Gayathri, V. M., & Supraja, P. (2020). Employee Attrition Prediction Using Machine Learning and Sentiment Analysis. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(5), 7550–7557. https://doi.org/10.30534/ijatcse/2020/91952020

Saleh, L., & Abu-Soud, S. (2021). Predicting Jordanian Job Satisfaction Using Artificial Neural Network and Decision Tree. *2021 11th International Conference on Advanced Computer Information Technologies, ACIT 2021 - Proceedings*, 735–738. https://doi.org/10.1109/ACIT52158.2021.9548364

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Seok, B. W., Wee, K. hoan, Park, J. young, Anil Kumar, D., & Reddy, N. S. (2021). Modeling the teacher job satisfaction by artificial neural networks. *Soft Computing*, *25*(17), 11803–11815. https://doi.org/10.1007/S00500-021-05958-0

Serrano-García, V., Ortega-Andeane, P., Reyes-Lagunes, I., & Riveros-Rosas, A. (2015). Traducción y Adaptación al Español del Cuestionario de Satisfacción Laboral para Profesores. *Acta de Investigación Psicológica*, *5*(3), 2112–2123. https://doi.org/10.1016/s2007-4719(16)30004-7

Silva, J., Zilberman, J., Romero, L., Pineda, O. B., & Herazo-Beltran, Y. (2020). Identification of Patterns of Fatal Injuries in Humans through Big Data. *Procedia Computer Science*, *170*, 893–898. https://doi.org/10.1016/j.procs.2020.03.114

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2018). Evaluation of machine learning

models for employee churn prediction. *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, 1016–1020. https://doi.org/10.1109/ICICI.2017.8365293

Skaalvik, E. M., & Skaalvik, S. (2011). Teacher job satisfaction and motivation to leave the teaching profession: Relations with school context, feeling of belonging, and emotional exhaustion. *Teaching and Teacher Education*, *27*(6), 1029–1038. https://doi.org/10.1016/j.tate.2011.04.001

Talingting, R. E. (2019). A data mining-driven model for job satisfaction prediction of school administrators in DepEd Surigao del Norte division. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(3), 556–560. https://doi.org/10.30534/IJATCSE/2019/34832019

Tomás, J. M., De Los Santos, S., & Fernández, I. (2019). Job satisfaction of the Dominican teacher: Labor background. *Revista Colombiana de Psicologia*, *28*(2), 63–76. https://doi.org/10.15446/rcp.v28n2.71675

Torres-Vásquez, M., Hernández-Torruco, J., Hernández-Ocaña, B., Chávez-Bosquez, O., Torres-Vásquez, M., Hernández-Torruco, J., Hernández-Ocaña, B., & Chávez-Bosquez, O. (2021). Impacto de los algoritmos de sobremuestreo en la clasificación de subtipos principales del síndrome de guillain-barré. *Ingenius. Revista de Ciencia y Tecnología* , *25*, 20–31. https://doi.org/10.17163/INGS.N25.2021.02

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543. https://doi.org/10.1016/J.ENBUILD.2017.11.039

Useche, L., & Mesa, D. (2006). Una introducción a la imputación de valores perdidos. *Terra*, *XXII*(31), 127–152.

Valcárcel, V. (2004). DATA MINING Y EL DESCUBRIMIENTO DEL CONOCIMIENTO (1)-9993 (electrónico). *Revista de La Facultad de Ingeniería Industrial*, *7*(2), 1810.

Van Loo, H. M., Bigdeli, T. B., Milaneschi, Y., Aggen, S. H., & Kendler, K. S. (2020). Data mining algorithm predicts a range of adverse outcomes in major depression. *Journal of Affective Disorders*, *276*, 945–953. https://doi.org/10.1016/j.jad.2020.07.098

Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/CAIT-2019-0001

Villa, A., Carrión, A., & Sozzi, A. (2016). Optimización del diseño de parámetros: Método Forest-Genetic univariante. *Publicaciones En Ciencias y Tecnología*, *10*(1), 12.

Yogesh, I., Suresh Kumar, K. R., Candrashekaran, N., Reddy, D., & Sampath, H. (2020). Predicting Job Satisfaction and Employee Turnover Using Machine Learning. *Journal of Computational and Theoretical Nanoscience*, *17*(9), 4092–4097. https://doi.org/10.1166/JCTN.2020.9024

Yoo, J. E., & Rho, M. (2020). Exploration of Predictors for Korean Teacher Job Satisfaction via a Machine Learning Technique, Group Mnet. *Frontiers in Psychology*, *11*, 441. https://doi.org/10.3389/fpsyg.2020.00441

Zembylas, M., & Papanastasiou, E. (2004). Job satisfaction among school teachers in Cyprus. *Journal of Educational Administration*, *42*(3), 357–374. https://doi.org/10.1108/09578230410534676

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (J. Bleiel & R. Roumeliotis (Eds.); First Edit).

ANEXOS

Anexo 1. Cuadro de matriz de consistencia

PROBLEMA		OBJETIVOS	HIPÓTESIS	VARIABLES	DIMENSIONES	ESCALA	METODOLOGÍA
GENERAL	¿Cuál de los modelos predictivos machine learning de regresión logística, árboles de decisión y clasificadores combinados permitirá obtener los mejores indicadores en la predicción de la satisfacción laboral de docentes de educación básica del Perú?	Determinar cuál de los modelos predictivos machine learning de regresión logística, árboles de decisión y clasificadores combinados permitirá obtener los mejores indicadores en la predicción de la satisfacción laboral de docentes de educación básica del Perú.	Los modelos predictivos machinen learning de regresión logística, árboles de decisión y clasificadores combinados obtendrán una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.	**Variable de estudio 1:** Modelo predictivo *machine learning*.	Modelo de regresión logística. Modelo de árboles de decisión (CART). Combinación de clasificadores (*Gradient Bosting*, *Random Forest*, *XG Boost*) **INDICADORES** -Área bajo la curva ROC (AUC). -Exactitud (Accuracy). -Sensibilidad (Recall-Tasa de verdaderos positivos). -Especificidad (Especifity-Tasa de verdaderos negativos). -Valor predictivo positivo. -Valor predictivo negativo. -Puntuación-F1 (F1-Score).	Escala para valor de AUC de acuerdo con Gironés et al. (2017): [0.5, 0.6 > *Test malo* [0.6, 0.7 > *Test regular* [0.75, 0.9 > *Test bueno* [0.9, 0.97 > *Test muy bueno* [0.97, 1.0 > *Test excelente*	***Tipo:*** El presente estudio corresponde a una investigación aplicada. (Arias, 2016). Bunge (1999) define este tipo de estudios como: "el campo de investigación en el que los problemas científicos con un posible sentido práctico se investigan con base en los descubrimientos de la ciencia básica (pura). Más que ser una investigación libre, tiene un objetivo, o mandato: de esa investigación se esperan eventualmente descubrimientos de interés práctico". (p. 277) ***Diseño de estudio:*** El presente estudio corresponde a un diseño no experimental de corte transversal. Estos estudios se realizan sin la manipulación deliberada de variables, donde solo observamos los fenómenos en su ambiente natural para analizarlos. Estos pueden ser corte transversal y longitudinal. Transversal, cuando la recolección de los datos se da en un solo memento. (Hernández Sampieri y Mendoza Torres, 2018, p. 176) ***Población y muestra***

ESPECÍFICOS	- ¿Cuáles son los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica? - ¿Cuál es el modelo de regresión logística para la satisfacción laboral de docentes de educación básica del Perú? - ¿Cómo es la predicción utilizando árboles de decisión para la satisfacción laboral de docentes de educación básica del Perú? - ¿Cómo es la predicción utilizando clasificadores combinados	-Identificar los atributos o variables más influyentes para el modelo de predicción de la satisfacción laboral de docentes de educación básica. -Establecer el modelo logístico para la satisfacción laboral de docentes de educación básica del Perú. -Predecir utilizando árboles de decisión la satisfacción laboral de docentes de educación básica del Perú. -Predecir utilizando clasificadores combinados la satisfacción laboral de docentes de	-El modelo construido con la técnica de regresión logística obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú. -El modelo construido con la técnica de árboles de decisión obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú. -El modelo construido con la técnica de clasificadores combinados	**Variable de estudio 2:** Satisfacción laboral docente.	**DIMENSIONES** Trabajo en sí mismo. Reconocimiento. Relación con los compañeros de trabajo. Pago (Salario). Condiciones de trabajo y Seguridad. **INDICADORES**	Insatisfecho Satisfecho	***Población:*** 470931 docentes de educación básica regular de instituciones educativas públicas y privadas del Perú en el 2018. ***Muestra:*** 15092 registros de la base de datos ENDO-2018. ***Técnicas e instrumentos:*** Algoritmos de aprendizaje automático supervisados. En la Encuesta Nacional a Docentes (ENDO-2018).

	para la satisfacción laboral de docentes de educación básica del Perú? - ¿Cuáles es el resultado de comparar la técnica de regresión logística respeto a las técnicas de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú?	educación básica del Perú. -Comparar las métricas de la técnica de regresión logística respeto a la técnica de árboles de decisión y clasificadores combinados en la predicción de la satisfacción laboral de docentes de educación básica del Perú.	Random Forest obtendrá una capacidad discriminativa buena en la predicción de la satisfacción laboral de docentes de educación básica del Perú.				

Anexo 2. Conjunto de datos de mediciones de métricas por modelo

	Modelo	Exactitud	Sensibilidad	VPP	F1-Score	AUC
0	Logistic Regression	0,7	0,725	0,690476	0,707317	0,81
1	Logistic Regression	0,75625	0,75	0,759494	0,754717	0,8267188
2	Logistic Regression	0,70625	0,725	0,698795	0,711656	0,7530469
3	Logistic Regression	0,66875	0,6	0,695652	0,644295	0,7873438
4	Logistic Regression	0,775	0,7875	0,768293	0,777778	0,8189844
5	Logistic Regression	0,725	0,7375	0,719512	0,728395	0,8057031
6	Logistic Regression	0,70625	0,7125	0,703704	0,708075	0,799375
7	Logistic Regression	0,74375	0,75	0,740741	0,745342	0,8040625
8	Logistic Regression	0,76875	0,725	0,794521	0,75817	0,8491406
9	Logistic Regression	0,71875	0,7375	0,710843	0,723926	0,8151563
10	Logistic Regression	0,79375	0,8	0,790123	0,795031	0,8441406
11	Logistic Regression	0,78125	0,7875	0,777778	0,782609	0,8503125
12	Logistic Regression	0,73125	0,7125	0,74026	0,726115	0,8225781
13	Logistic Regression	0,75625	0,75	0,759494	0,754717	0,87
14	Logistic Regression	0,8	0,8375	0,77907	0,807229	0,8753125
15	Logistic Regression	0,725	0,6625	0,757143	0,706667	0,8015625
16	Logistic Regression	0,775	0,7875	0,768293	0,777778	0,858125
17	Logistic Regression	0,74375	0,775	0,729412	0,751515	0,8270313
18	Logistic Regression	0,76875	0,775	0,765432	0,770186	0,8504688
19	Logistic Regression	0,73125	0,7875	0,707865	0,745562	0,7994531
20	Logistic Regression	0,74375	0,75	0,740741	0,745342	0,8103906
21	Logistic Regression	0,7125	0,7375	0,702381	0,719512	0,8234375
22	Logistic Regression	0,75625	0,8	0,735632	0,766467	0,8407813
23	Logistic Regression	0,7875	0,775	0,794872	0,78481	0,888125
24	Logistic Regression	0,75625	0,875	0,707071	0,782123	0,8285156
25	Logistic Regression	0,71875	0,725	0,716049	0,720497	0,8145313
26	Logistic Regression	0,7625	0,8125	0,738636	0,77381	0,8282813
27	Logistic Regression	0,7375	0,8125	0,706522	0,755814	0,8254688
28	Logistic Regression	0,725	0,725	0,725	0,725	0,7658594
29	Logistic Regression	0,75625	0,8125	0,730337	0,769231	0,845625
30	Logistic Regression	0,74375	0,7875	0,724138	0,754491	0,805625
31	Logistic Regression	0,725	0,7625	0,709302	0,73494	0,826875
32	Logistic Regression	0,775	0,7625	0,782051	0,772152	0,8459375
33	Logistic Regression	0,70625	0,7	0,708861	0,704403	0,7932813
34	Logistic Regression	0,6875	0,675	0,692308	0,683544	0,7926563
35	Logistic Regression	0,7375	0,8125	0,706522	0,755814	0,8103906
36	Logistic Regression	0,73125	0,7375	0,728395	0,732919	0,8095313
37	Logistic Regression	0,74375	0,775	0,729412	0,751515	0,8313281
38	Logistic Regression	0,71875	0,7625	0,701149	0,730539	0,8057813
39	Logistic Regression	0,75625	0,7	0,788732	0,741722	0,8394531
40	Logistic Regression	0,698113	0,683544	0,701299	0,692308	0,7871044
41	Logistic Regression	0,761006	0,810127	0,735632	0,771084	0,840981
42	Logistic Regression	0,805031	0,848101	0,77907	0,812121	0,8863924
43	Logistic Regression	0,761006	0,78481	0,746988	0,765432	0,8554589

44	Logistic Regression	0,72956	0,734177	0,725	0,72956	0,7985759
45	Logistic Regression	0,685535	0,632911	0,704225	0,666667	0,7933544
46	Logistic Regression	0,754717	0,835443	0,717391	0,77193	0,8371044
47	Logistic Regression	0,704403	0,670886	0,716216	0,69281	0,8022943
48	Logistic Regression	0,716981	0,759494	0,697674	0,727273	0,7820411
49	Logistic Regression	0,72956	0,683544	0,75	0,715232	0,8126582
50	Logistic Regression	0,704403	0,683544	0,710526	0,696774	0,7887658
51	Logistic Regression	0,742138	0,78481	0,72093	0,751515	0,7994462
52	Logistic Regression	0,685535	0,746835	0,662921	0,702381	0,7492089
53	Logistic Regression	0,72956	0,696203	0,743243	0,718954	0,841693
54	Logistic Regression	0,716981	0,658228	0,742857	0,697987	0,8096519
55	Logistic Regression	0,72327	0,746835	0,710843	0,728395	0,7885285
56	Logistic Regression	0,798742	0,822785	0,783133	0,802469	0,8580696
57	Logistic Regression	0,779874	0,772152	0,782051	0,77707	0,8468354
58	Logistic Regression	0,805031	0,848101	0,77907	0,812121	0,8636076
59	Logistic Regression	0,710692	0,721519	0,703704	0,7125	0,7943829
60	Logistic Regression	0,748428	0,772152	0,73494	0,753086	0,8264241
61	Logistic Regression	0,647799	0,620253	0,653333	0,636364	0,7329905
62	Logistic Regression	0,742138	0,708861	0,756757	0,732026	0,8355222
63	Logistic Regression	0,735849	0,78481	0,712644	0,746988	0,7884494
64	Logistic Regression	0,767296	0,734177	0,783784	0,75817	0,860443
65	Logistic Regression	0,72956	0,746835	0,719512	0,732919	0,7892405
66	Logistic Regression	0,805031	0,835443	0,785714	0,809816	0,8917722
67	Logistic Regression	0,773585	0,797468	0,759036	0,777778	0,8345728
68	Logistic Regression	0,742138	0,721519	0,75	0,735484	0,8381329
69	Logistic Regression	0,786164	0,78481	0,78481	0,78481	0,8626582
70	Logistic Regression	0,761006	0,8	0,744186	0,771084	0,8151108
71	Logistic Regression	0,716981	0,775	0,696629	0,733728	0,8011076
72	Logistic Regression	0,654088	0,7375	0,634409	0,682081	0,7426424
73	Logistic Regression	0,742138	0,7875	0,724138	0,754491	0,8015823
74	Logistic Regression	0,685535	0,7	0,682927	0,691358	0,7738133
75	Logistic Regression	0,754717	0,775	0,746988	0,760736	0,8359177
76	Logistic Regression	0,748428	0,7875	0,732558	0,759036	0,7973101
77	Logistic Regression	0,716981	0,65	0,753623	0,697987	0,8066456
78	Logistic Regression	0,704403	0,7375	0,694118	0,715152	0,8045886
79	Logistic Regression	0,735849	0,7375	0,7375	0,7375	0,8183544
80	Logistic Regression	0,742138	0,7625	0,73494	0,748466	0,8455696
81	Logistic Regression	0,691824	0,6875	0,696203	0,691824	0,8122627
82	Logistic Regression	0,72327	0,7375	0,719512	0,728395	0,8206487
83	Logistic Regression	0,748428	0,7125	0,77027	0,74026	0,8314873
84	Logistic Regression	0,742138	0,7875	0,724138	0,754491	0,8246835
85	Logistic Regression	0,72327	0,7375	0,719512	0,728395	0,8049051
86	Logistic Regression	0,735849	0,7125	0,75	0,730769	0,814557
87	Logistic Regression	0,710692	0,6625	0,736111	0,697368	0,8146361
88	Logistic Regression	0,704403	0,7	0,708861	0,704403	0,7946203
89	Logistic Regression	0,754717	0,775	0,746988	0,760736	0,822231

90	Logistic Regression	0,761006	0,7625	0,7625	0,7625	0,8772943
91	Logistic Regression	0,691824	0,6625	0,706667	0,683871	0,7911392
92	Logistic Regression	0,716981	0,7	0,727273	0,713376	0,7974684
93	Logistic Regression	0,773585	0,7	0,823529	0,756757	0,8665348
94	Logistic Regression	0,72956	0,75	0,722892	0,736196	0,8159019
95	Logistic Regression	0,716981	0,6875	0,733333	0,709677	0,8242089
96	Logistic Regression	0,716981	0,775	0,696629	0,733728	0,828481
97	Logistic Regression	0,767296	0,8375	0,736264	0,783626	0,8613924
98	Logistic Regression	0,698113	0,7625	0,677778	0,717647	0,7855222
99	Logistic Regression	0,685535	0,7	0,682927	0,691358	0,7864715
100	Gradient Boosting	0,80625	0,875	0,769231	0,818713	0,87875
101	Gradient Boosting	0,80625	0,8125	0,802469	0,807453	0,8911719
102	Gradient Boosting	0,76875	0,8125	0,747126	0,778443	0,8521875
103	Gradient Boosting	0,79375	0,7625	0,813333	0,787097	0,8754688
104	Gradient Boosting	0,8125	0,8375	0,797619	0,817073	0,8780469
105	Gradient Boosting	0,78125	0,8	0,771084	0,785276	0,8780469
106	Gradient Boosting	0,825	0,875	0,795455	0,833333	0,905
107	Gradient Boosting	0,76875	0,8	0,752941	0,775758	0,845625
108	Gradient Boosting	0,8375	0,8	0,864865	0,831169	0,9110156
109	Gradient Boosting	0,81875	0,8625	0,793103	0,826347	0,9000781
110	Gradient Boosting	0,83125	0,85	0,819277	0,834356	0,9044531
111	Gradient Boosting	0,85625	0,875	0,843373	0,858896	0,9220313
112	Gradient Boosting	0,78125	0,75	0,8	0,774194	0,8854688
113	Gradient Boosting	0,8	0,825	0,785714	0,804878	0,89125
114	Gradient Boosting	0,8375	0,8625	0,821429	0,841463	0,9203906
115	Gradient Boosting	0,8375	0,85	0,829268	0,839506	0,8907813
116	Gradient Boosting	0,8375	0,85	0,829268	0,839506	0,9142969
117	Gradient Boosting	0,85625	0,875	0,843373	0,858896	0,91875
118	Gradient Boosting	0,84375	0,8625	0,831325	0,846626	0,919375
119	Gradient Boosting	0,7625	0,775	0,756098	0,765432	0,854375
120	Gradient Boosting	0,8125	0,8125	0,8125	0,8125	0,8969531
121	Gradient Boosting	0,775	0,825	0,75	0,785714	0,8809375
122	Gradient Boosting	0,84375	0,8875	0,816092	0,850299	0,9189063
123	Gradient Boosting	0,85625	0,8875	0,835294	0,860606	0,94125
124	Gradient Boosting	0,7875	0,875	0,744681	0,804598	0,8744531
125	Gradient Boosting	0,79375	0,8125	0,783133	0,797546	0,8892188
126	Gradient Boosting	0,80625	0,875	0,769231	0,818713	0,8875781
127	Gradient Boosting	0,8	0,85	0,772727	0,809524	0,8967188
128	Gradient Boosting	0,79375	0,8125	0,783133	0,797546	0,8700781
129	Gradient Boosting	0,79375	0,8625	0,758242	0,807018	0,9048438
130	Gradient Boosting	0,8125	0,8875	0,771739	0,825581	0,886875
131	Gradient Boosting	0,81875	0,8875	0,78022	0,830409	0,9116406
132	Gradient Boosting	0,84375	0,8625	0,831325	0,846626	0,9089063
133	Gradient Boosting	0,78125	0,8125	0,764706	0,787879	0,8709375
134	Gradient Boosting	0,83125	0,825	0,835443	0,830189	0,8939063
135	Gradient Boosting	0,8	0,825	0,785714	0,804878	0,9047656

136	Gradient Boosting	0,78125	0,8125	0,764706	0,787879	0,8615625
137	Gradient Boosting	0,8375	0,825	0,846154	0,835443	0,9017188
138	Gradient Boosting	0,8125	0,85	0,790698	0,819277	0,8788281
139	Gradient Boosting	0,83125	0,825	0,835443	0,830189	0,9057031
140	Gradient Boosting	0,805031	0,810127	0,8	0,805031	0,8831487
141	Gradient Boosting	0,792453	0,835443	0,767442	0,8	0,8868671
142	Gradient Boosting	0,880503	0,936709	0,840909	0,886228	0,9346519
143	Gradient Boosting	0,811321	0,860759	0,781609	0,819277	0,9124209
144	Gradient Boosting	0,805031	0,873418	0,766667	0,816568	0,8845728
145	Gradient Boosting	0,786164	0,797468	0,777778	0,7875	0,8668513
146	Gradient Boosting	0,811321	0,898734	0,763441	0,825581	0,9157437
147	Gradient Boosting	0,773585	0,746835	0,786667	0,766234	0,8633703
148	Gradient Boosting	0,81761	0,898734	0,771739	0,830409	0,8760285
149	Gradient Boosting	0,786164	0,772152	0,792208	0,782051	0,8685918
150	Gradient Boosting	0,786164	0,772152	0,792208	0,782051	0,8626582
151	Gradient Boosting	0,798742	0,860759	0,764045	0,809524	0,8906646
152	Gradient Boosting	0,735849	0,78481	0,712644	0,746988	0,8389241
153	Gradient Boosting	0,767296	0,759494	0,769231	0,764331	0,9037184
154	Gradient Boosting	0,773585	0,78481	0,765432	0,775	0,8696203
155	Gradient Boosting	0,792453	0,835443	0,767442	0,8	0,8820411
156	Gradient Boosting	0,880503	0,962025	0,826087	0,888889	0,9457278
157	Gradient Boosting	0,842767	0,860759	0,829268	0,84472	0,9050633
158	Gradient Boosting	0,842767	0,848101	0,8375	0,842767	0,9053797
159	Gradient Boosting	0,773585	0,810127	0,752941	0,780488	0,8649525
160	Gradient Boosting	0,805031	0,848101	0,77907	0,812121	0,8825949
161	Gradient Boosting	0,735849	0,721519	0,74026	0,730769	0,8120253
162	Gradient Boosting	0,867925	0,898734	0,845238	0,871166	0,9280063
163	Gradient Boosting	0,805031	0,848101	0,77907	0,812121	0,8431962
164	Gradient Boosting	0,849057	0,873418	0,831325	0,851852	0,944462
165	Gradient Boosting	0,805031	0,810127	0,8	0,805031	0,8832278
166	Gradient Boosting	0,893082	0,974684	0,836957	0,900585	0,9713608
167	Gradient Boosting	0,849057	0,886076	0,823529	0,853659	0,9034019
168	Gradient Boosting	0,836478	0,860759	0,819277	0,839506	0,9150316
169	Gradient Boosting	0,849057	0,860759	0,839506	0,85	0,9366297
170	Gradient Boosting	0,792453	0,8125	0,783133	0,797546	0,8602848
171	Gradient Boosting	0,811321	0,8625	0,784091	0,821429	0,8704114
172	Gradient Boosting	0,72956	0,825	0,694737	0,754286	0,8367089
173	Gradient Boosting	0,805031	0,85	0,781609	0,814371	0,8848101
174	Gradient Boosting	0,754717	0,7875	0,741176	0,763636	0,8660601
175	Gradient Boosting	0,849057	0,875	0,833333	0,853659	0,9131329
176	Gradient Boosting	0,798742	0,8125	0,792683	0,802469	0,8780063
177	Gradient Boosting	0,81761	0,8375	0,807229	0,822086	0,8914557
178	Gradient Boosting	0,798742	0,8125	0,792683	0,802469	0,8928797
179	Gradient Boosting	0,805031	0,8125	0,802469	0,807453	0,8982595
180	Gradient Boosting	0,792453	0,8375	0,770115	0,802395	0,8820411
181	Gradient Boosting	0,830189	0,875	0,804598	0,838323	0,9076741

182	Gradient Boosting	0,767296	0,8125	0,747126	0,778443	0,8743671
183	Gradient Boosting	0,798742	0,8375	0,77907	0,807229	0,8757911
184	Gradient Boosting	0,811321	0,875	0,777778	0,823529	0,9151899
185	Gradient Boosting	0,811321	0,875	0,777778	0,823529	0,9009494
186	Gradient Boosting	0,867925	0,9	0,847059	0,872727	0,9134494
187	Gradient Boosting	0,805031	0,825	0,795181	0,809816	0,8920095
188	Gradient Boosting	0,761006	0,7875	0,75	0,768293	0,8767405
189	Gradient Boosting	0,830189	0,8375	0,82716	0,832298	0,9159019
190	Gradient Boosting	0,861635	0,875	0,853659	0,864198	0,9263449
191	Gradient Boosting	0,811321	0,8625	0,784091	0,821429	0,8773734
192	Gradient Boosting	0,779874	0,8125	0,764706	0,787879	0,8693829
193	Gradient Boosting	0,836478	0,775	0,885714	0,826667	0,9258703
194	Gradient Boosting	0,81761	0,825	0,814815	0,819876	0,9016614
195	Gradient Boosting	0,849057	0,875	0,833333	0,853659	0,9218354
196	Gradient Boosting	0,830189	0,8875	0,797753	0,840237	0,9120253
197	Gradient Boosting	0,867925	0,8875	0,855422	0,871166	0,9317247
198	Gradient Boosting	0,754717	0,8125	0,730337	0,769231	0,8577532
199	Gradient Boosting	0,761006	0,8125	0,738636	0,77381	0,8755538
200	Random Forest	0,75	0,85	0,708333	0,772727	0,8214844
201	Random Forest	0,7625	0,8	0,744186	0,771084	0,8430469
202	Random Forest	0,73125	0,7875	0,707865	0,745562	0,7976563
203	Random Forest	0,7125	0,6625	0,736111	0,697368	0,8197656
204	Random Forest	0,75625	0,825	0,725275	0,77193	0,8436719
205	Random Forest	0,725	0,8375	0,683673	0,752809	0,8385156
206	Random Forest	0,75	0,85	0,708333	0,772727	0,8421875
207	Random Forest	0,7375	0,7875	0,715909	0,75	0,8097656
208	Random Forest	0,7875	0,7625	0,802632	0,782051	0,8719531
209	Random Forest	0,78125	0,85	0,747253	0,795322	0,8525781
210	Random Forest	0,80625	0,8625	0,775281	0,816568	0,8489063
211	Random Forest	0,8	0,8625	0,766667	0,811765	0,8975781
212	Random Forest	0,74375	0,7375	0,746835	0,742138	0,8384375
213	Random Forest	0,775	0,875	0,729167	0,795455	0,8723438
214	Random Forest	0,80625	0,8625	0,775281	0,816568	0,906875
215	Random Forest	0,775	0,775	0,775	0,775	0,8230469
216	Random Forest	0,8125	0,8375	0,797619	0,817073	0,8821094
217	Random Forest	0,78125	0,85	0,747253	0,795322	0,851875
218	Random Forest	0,78125	0,8125	0,764706	0,787879	0,875625
219	Random Forest	0,725	0,7875	0,7	0,741176	0,8214844
220	Random Forest	0,775	0,8	0,761905	0,780488	0,8447656
221	Random Forest	0,75	0,8375	0,712766	0,770115	0,8474219
222	Random Forest	0,8	0,875	0,76087	0,813953	0,8778906
223	Random Forest	0,8375	0,8875	0,806818	0,845238	0,919375
224	Random Forest	0,73125	0,85	0,686869	0,759777	0,8303906
225	Random Forest	0,73125	0,7375	0,728395	0,732919	0,8508594
226	Random Forest	0,75	0,825	0,717391	0,767442	0,8467188
227	Random Forest	0,75625	0,875	0,707071	0,782123	0,8528906

228	Random Forest	0,75	0,775	0,738095	0,756098	0,7946875
229	Random Forest	0,75625	0,8125	0,730337	0,769231	0,8407813
230	Random Forest	0,78125	0,85	0,747253	0,795322	0,8455469
231	Random Forest	0,74375	0,8	0,719101	0,757396	0,8532813
232	Random Forest	0,8	0,875	0,76087	0,813953	0,8642188
233	Random Forest	0,7375	0,775	0,72093	0,746988	0,8152344
234	Random Forest	0,7375	0,75	0,731707	0,740741	0,8339063
235	Random Forest	0,75625	0,825	0,725275	0,77193	0,8263281
236	Random Forest	0,75	0,8	0,727273	0,761905	0,8225
237	Random Forest	0,7875	0,825	0,767442	0,795181	0,8600781
238	Random Forest	0,7375	0,7875	0,715909	0,75	0,8310938
239	Random Forest	0,75	0,7625	0,743902	0,753086	0,85125
240	Random Forest	0,754717	0,746835	0,75641	0,751592	0,8136867
241	Random Forest	0,773585	0,848101	0,736264	0,788235	0,8525316
242	Random Forest	0,836478	0,911392	0,791209	0,847059	0,9048259
243	Random Forest	0,798742	0,860759	0,764045	0,809524	0,8772943
244	Random Forest	0,779874	0,848101	0,744444	0,792899	0,8344937
245	Random Forest	0,72956	0,78481	0,704545	0,742515	0,8363924
246	Random Forest	0,767296	0,924051	0,701923	0,797814	0,8686709
247	Random Forest	0,710692	0,721519	0,703704	0,7125	0,8218354
248	Random Forest	0,779874	0,886076	0,729167	0,8	0,8300633
249	Random Forest	0,742138	0,759494	0,731707	0,745342	0,8084652
250	Random Forest	0,735849	0,708861	0,746667	0,727273	0,8113133
251	Random Forest	0,761006	0,848101	0,72043	0,77907	0,8346519
252	Random Forest	0,685535	0,78481	0,652632	0,712644	0,774288
253	Random Forest	0,754717	0,759494	0,75	0,754717	0,8466772
254	Random Forest	0,742138	0,759494	0,731707	0,745342	0,8268987
255	Random Forest	0,72327	0,822785	0,684211	0,747126	0,8298259
256	Random Forest	0,811321	0,936709	0,747475	0,831461	0,8961234
257	Random Forest	0,811321	0,822785	0,802469	0,8125	0,8641614
258	Random Forest	0,754717	0,810127	0,727273	0,766467	0,8678006
259	Random Forest	0,716981	0,759494	0,697674	0,727273	0,8204905
260	Random Forest	0,792453	0,848101	0,761364	0,802395	0,8682753
261	Random Forest	0,679245	0,708861	0,666667	0,687117	0,7417722
262	Random Forest	0,761006	0,78481	0,746988	0,765432	0,8704905
263	Random Forest	0,710692	0,759494	0,689655	0,722892	0,8060127
264	Random Forest	0,811321	0,822785	0,802469	0,8125	0,9155063
265	Random Forest	0,735849	0,746835	0,728395	0,7375	0,8303006
266	Random Forest	0,836478	0,911392	0,791209	0,847059	0,9371044
267	Random Forest	0,767296	0,822785	0,738636	0,778443	0,847231
268	Random Forest	0,761006	0,78481	0,746988	0,765432	0,8605222
269	Random Forest	0,767296	0,848101	0,728261	0,783626	0,8897943
270	Random Forest	0,767296	0,8	0,752941	0,775758	0,8120253
271	Random Forest	0,792453	0,8625	0,758242	0,807018	0,8338608
272	Random Forest	0,666667	0,8	0,633663	0,707182	0,7623418
273	Random Forest	0,72327	0,775	0,704545	0,738095	0,8257911

274	Random Forest	0,742138	0,8	0,719101	0,757396	0,809731
275	Random Forest	0,754717	0,85	0,715789	0,777143	0,8518196
276	Random Forest	0,754717	0,8	0,735632	0,766467	0,8394778
277	Random Forest	0,792453	0,825	0,776471	0,8	0,8556171
278	Random Forest	0,754717	0,775	0,746988	0,760736	0,8524525
279	Random Forest	0,773585	0,8125	0,755814	0,783133	0,8428006
280	Random Forest	0,767296	0,825	0,741573	0,781065	0,8636867
281	Random Forest	0,773585	0,85	0,73913	0,790698	0,8540348
282	Random Forest	0,72956	0,7875	0,707865	0,745562	0,8431171
283	Random Forest	0,754717	0,7625	0,753086	0,757764	0,8464399
284	Random Forest	0,805031	0,875	0,769231	0,818713	0,8911392
285	Random Forest	0,754717	0,8125	0,730337	0,769231	0,8526899
286	Random Forest	0,773585	0,8	0,761905	0,780488	0,844462
287	Random Forest	0,761006	0,75	0,769231	0,759494	0,8443038
288	Random Forest	0,716981	0,7375	0,710843	0,723926	0,8067247
289	Random Forest	0,792453	0,8375	0,770115	0,802395	0,8810918
290	Random Forest	0,830189	0,8875	0,797753	0,840237	0,8924842
291	Random Forest	0,742138	0,7875	0,724138	0,754491	0,8332278
292	Random Forest	0,735849	0,7875	0,715909	0,75	0,7987342
293	Random Forest	0,805031	0,75	0,84507	0,794702	0,8984177
294	Random Forest	0,761006	0,7875	0,75	0,768293	0,8416139
295	Random Forest	0,798742	0,8	0,8	0,8	0,8734968
296	Random Forest	0,748428	0,8125	0,722222	0,764706	0,8567247
297	Random Forest	0,823899	0,8625	0,802326	0,831325	0,8874209
298	Random Forest	0,754717	0,875	0,707071	0,782123	0,8004747
299	Random Forest	0,710692	0,775	0,688889	0,729412	0,8085443
300	XGBoost	0,80625	0,85	0,781609	0,814371	0,8864063
301	XGBoost	0,8125	0,8	0,820513	0,810127	0,904375
302	XGBoost	0,8	0,8625	0,766667	0,811765	0,88
303	XGBoost	0,78125	0,7375	0,808219	0,771242	0,8798438
304	XGBoost	0,83125	0,8375	0,82716	0,832298	0,8978906
305	XGBoost	0,8125	0,825	0,804878	0,814815	0,8825781
306	XGBoost	0,84375	0,8625	0,831325	0,846626	0,9140625
307	XGBoost	0,79375	0,8	0,790123	0,795031	0,8745313
308	XGBoost	0,81875	0,8125	0,822785	0,81761	0,9216406
309	XGBoost	0,84375	0,875	0,823529	0,848485	0,9060938
310	XGBoost	0,85625	0,9	0,827586	0,862275	0,9169531
311	XGBoost	0,8625	0,875	0,853659	0,864198	0,93125
312	XGBoost	0,79375	0,7625	0,813333	0,787097	0,8847656
313	XGBoost	0,81875	0,85	0,8	0,824242	0,8790625
314	XGBoost	0,85625	0,875	0,843373	0,858896	0,940625
315	XGBoost	0,8375	0,8125	0,855263	0,833333	0,9076563
316	XGBoost	0,85	0,8375	0,858974	0,848101	0,9247656
317	XGBoost	0,875	0,9	0,857143	0,878049	0,9283594
318	XGBoost	0,8625	0,8875	0,845238	0,865854	0,9401563
319	XGBoost	0,7625	0,7875	0,75	0,768293	0,8685156

320	XGBoost	0,825	0,825	0,825	0,825	0,9060156
321	XGBoost	0,8125	0,85	0,790698	0,819277	0,8979688
322	XGBoost	0,8375	0,8625	0,821429	0,841463	0,9260938
323	XGBoost	0,8375	0,8875	0,806818	0,845238	0,9314063
324	XGBoost	0,775	0,85	0,73913	0,790698	0,8660156
325	XGBoost	0,80625	0,8125	0,802469	0,807453	0,9120313
326	XGBoost	0,83125	0,9	0,791209	0,842105	0,9179688
327	XGBoost	0,79375	0,875	0,752688	0,809249	0,9092188
328	XGBoost	0,83125	0,8625	0,811765	0,836364	0,8867969
329	XGBoost	0,83125	0,9	0,791209	0,842105	0,9221875
330	XGBoost	0,8375	0,875	0,813953	0,843373	0,8901563
331	XGBoost	0,81875	0,8625	0,793103	0,826347	0,9178125
332	XGBoost	0,84375	0,9	0,808989	0,852071	0,9173438
333	XGBoost	0,81875	0,85	0,8	0,824242	0,9034375
334	XGBoost	0,85625	0,8625	0,851852	0,857143	0,920625
335	XGBoost	0,84375	0,8875	0,816092	0,850299	0,9271094
336	XGBoost	0,8125	0,825	0,804878	0,814815	0,8954688
337	XGBoost	0,8625	0,8875	0,845238	0,865854	0,9100781
338	XGBoost	0,825	0,875	0,795455	0,833333	0,9028125
339	XGBoost	0,8375	0,8375	0,8375	0,8375	0,9322656
340	XGBoost	0,798742	0,835443	0,776471	0,804878	0,8717563
341	XGBoost	0,786164	0,822785	0,764706	0,792683	0,8981013
342	XGBoost	0,886792	0,936709	0,850575	0,891566	0,9474684
343	XGBoost	0,849057	0,873418	0,831325	0,851852	0,9173259
344	XGBoost	0,823899	0,873418	0,793103	0,831325	0,8977848
345	XGBoost	0,798742	0,822785	0,783133	0,802469	0,8794304
346	XGBoost	0,830189	0,936709	0,770833	0,845714	0,9320411
347	XGBoost	0,767296	0,746835	0,776316	0,76129	0,8587816
348	XGBoost	0,830189	0,898734	0,788889	0,840237	0,8926424
349	XGBoost	0,811321	0,797468	0,818182	0,807692	0,8952532
350	XGBoost	0,842767	0,822785	0,855263	0,83871	0,9006329
351	XGBoost	0,823899	0,886076	0,786517	0,833333	0,9084652
352	XGBoost	0,767296	0,835443	0,733333	0,781065	0,8768987
353	XGBoost	0,81761	0,78481	0,837838	0,810458	0,9246044
354	XGBoost	0,849057	0,848101	0,848101	0,848101	0,9023734
355	XGBoost	0,842767	0,886076	0,813953	0,848485	0,8965981
356	XGBoost	0,90566	0,987342	0,847826	0,912281	0,9582278
357	XGBoost	0,855346	0,873418	0,841463	0,857143	0,9060127
358	XGBoost	0,849057	0,835443	0,857143	0,846154	0,9281646
359	XGBoost	0,798742	0,835443	0,776471	0,804878	0,8962816
360	XGBoost	0,792453	0,835443	0,767442	0,8	0,8759494
361	XGBoost	0,761006	0,746835	0,766234	0,75641	0,8649525
362	XGBoost	0,867925	0,886076	0,853659	0,869565	0,9364715
363	XGBoost	0,842767	0,886076	0,813953	0,848485	0,8838608
364	XGBoost	0,874214	0,873418	0,873418	0,873418	0,9484177
365	XGBoost	0,836478	0,860759	0,819277	0,839506	0,8905063

366	XGBoost	0,886792	0,987342	0,821053	0,896552	0,9610759
367	XGBoost	0,861635	0,873418	0,851852	0,8625	0,9198576
368	XGBoost	0,849057	0,860759	0,839506	0,85	0,9061709
369	XGBoost	0,855346	0,873418	0,841463	0,857143	0,9297468
370	XGBoost	0,811321	0,8375	0,797619	0,817073	0,8670095
371	XGBoost	0,811321	0,8625	0,784091	0,821429	0,8775316
372	XGBoost	0,748428	0,85	0,708333	0,772727	0,8358386
373	XGBoost	0,805031	0,8375	0,788235	0,812121	0,9248418
374	XGBoost	0,830189	0,85	0,819277	0,834356	0,8951741
375	XGBoost	0,830189	0,8625	0,811765	0,836364	0,9243671
376	XGBoost	0,81761	0,85	0,8	0,824242	0,9025316
377	XGBoost	0,855346	0,875	0,843373	0,858896	0,9099684
378	XGBoost	0,792453	0,8	0,790123	0,795031	0,8933544
379	XGBoost	0,836478	0,825	0,846154	0,835443	0,915981
380	XGBoost	0,805031	0,8375	0,788235	0,812121	0,8974684
381	XGBoost	0,830189	0,8875	0,797753	0,840237	0,9284019
382	XGBoost	0,786164	0,825	0,767442	0,795181	0,8964399
383	XGBoost	0,830189	0,875	0,804598	0,838323	0,8625
384	XGBoost	0,81761	0,875	0,786517	0,828402	0,9082278
385	XGBoost	0,842767	0,8875	0,816092	0,850299	0,9243671
386	XGBoost	0,874214	0,8875	0,865854	0,876543	0,9286392
387	XGBoost	0,842767	0,8875	0,816092	0,850299	0,9215981
388	XGBoost	0,823899	0,8625	0,802326	0,831325	0,9041139
389	XGBoost	0,823899	0,85	0,809524	0,829268	0,920807
390	XGBoost	0,874214	0,9	0,857143	0,878049	0,9359968
391	XGBoost	0,805031	0,875	0,769231	0,818713	0,8984177
392	XGBoost	0,811321	0,875	0,777778	0,823529	0,8957278
393	XGBoost	0,874214	0,85	0,894737	0,871795	0,9404272
394	XGBoost	0,842767	0,8875	0,816092	0,850299	0,9136867
395	XGBoost	0,867925	0,9	0,847059	0,872727	0,9226266
396	XGBoost	0,823899	0,9125	0,776596	0,83908	0,9253165
397	XGBoost	0,861635	0,8875	0,845238	0,865854	0,9375
398	XGBoost	0,773585	0,8	0,761905	0,780488	0,8728639
399	XGBoost	0,792453	0,825	0,776471	0,8	0,8984968
400	Decision Trees-CART	0,76875	0,7875	0,759036	0,773006	0,8377344
401	Decision Trees-CART	0,7125	0,625	0,757576	0,684932	0,811875
402	Decision Trees-CART	0,69375	0,6375	0,71831	0,675497	0,7555469
403	Decision Trees-CART	0,725	0,5875	0,810345	0,681159	0,8117969
404	Decision Trees-CART	0,8	0,75	0,833333	0,789474	0,845
405	Decision Trees-CART	0,7375	0,6875	0,763889	0,723684	0,8209375
406	Decision Trees-CART	0,76875	0,7125	0,802817	0,754967	0,8271094
407	Decision Trees-CART	0,7125	0,675	0,72973	0,701299	0,7751563

408	Decision Trees-CART	0,75625	0,6625	0,815385	0,731034	0,8564844
409	Decision Trees-CART	0,775	0,7375	0,797297	0,766234	0,845
410	Decision Trees-CART	0,76875	0,7375	0,786667	0,76129	0,8105469
411	Decision Trees-CART	0,80625	0,7875	0,818182	0,802548	0,8807031
412	Decision Trees-CART	0,7375	0,65	0,787879	0,712329	0,8328125
413	Decision Trees-CART	0,74375	0,7	0,767123	0,732026	0,8278906
414	Decision Trees-CART	0,7625	0,7	0,8	0,746667	0,8875
415	Decision Trees-CART	0,7375	0,65	0,787879	0,712329	0,8103125
416	Decision Trees-CART	0,8	0,7625	0,824324	0,792208	0,8866406
417	Decision Trees-CART	0,7625	0,7125	0,791667	0,75	0,8340625
418	Decision Trees-CART	0,75625	0,725	0,773333	0,748387	0,8590625
419	Decision Trees-CART	0,7625	0,7125	0,791667	0,75	0,8330469
420	Decision Trees-CART	0,73125	0,6375	0,784615	0,703448	0,8170313
421	Decision Trees-CART	0,74375	0,6875	0,774648	0,728477	0,8252344
422	Decision Trees-CART	0,7625	0,725	0,783784	0,753247	0,86875
423	Decision Trees-CART	0,825	0,775	0,861111	0,815789	0,8952344
424	Decision Trees-CART	0,69375	0,725	0,682353	0,70303	0,7835938
425	Decision Trees-CART	0,75	0,6875	0,785714	0,733333	0,8509375
426	Decision Trees-CART	0,75625	0,725	0,773333	0,748387	0,8342969
427	Decision Trees-CART	0,71875	0,7375	0,710843	0,723926	0,8371094
428	Decision Trees-CART	0,70625	0,625	0,746269	0,680272	0,7728125
429	Decision Trees-CART	0,75625	0,725	0,773333	0,748387	0,83
430	Decision Trees-CART	0,71875	0,65	0,753623	0,697987	0,8185938
431	Decision Trees-CART	0,75	0,75	0,75	0,75	0,8186719
432	Decision Trees-CART	0,7875	0,75	0,810811	0,779221	0,8428125
433	Decision Trees-CART	0,75	0,7125	0,77027	0,74026	0,8180469
434	Decision Trees-CART	0,75	0,7	0,777778	0,736842	0,8277344
435	Decision Trees-CART	0,7375	0,7125	0,75	0,730769	0,8025
436	Decision Trees-CART	0,75	0,6875	0,785714	0,733333	0,8316406

437	Decision Trees-CART	0,75625	0,6875	0,797101	0,738255	0,8354688
438	Decision Trees-CART	0,7125	0,65	0,742857	0,693333	0,8069531
439	Decision Trees-CART	0,75625	0,6625	0,815385	0,731034	0,85
440	Decision Trees-CART	0,72327	0,620253	0,777778	0,690141	0,8
441	Decision Trees-CART	0,748428	0,734177	0,753247	0,74359	0,8441456
442	Decision Trees-CART	0,805031	0,78481	0,815789	0,8	0,8852057
443	Decision Trees-CART	0,735849	0,721519	0,74026	0,730769	0,8049051
444	Decision Trees-CART	0,72327	0,632911	0,769231	0,694444	0,8014241
445	Decision Trees-CART	0,710692	0,632911	0,746269	0,684932	0,7912975
446	Decision Trees-CART	0,798742	0,797468	0,797468	0,797468	0,8490506
447	Decision Trees-CART	0,72956	0,632911	0,78125	0,699301	0,8133703
448	Decision Trees-CART	0,704403	0,658228	0,722222	0,688742	0,7977057
449	Decision Trees-CART	0,704403	0,620253	0,742424	0,675862	0,7841772
450	Decision Trees-CART	0,716981	0,632911	0,757576	0,689655	0,7989715
451	Decision Trees-CART	0,773585	0,772152	0,772152	0,772152	0,8411392
452	Decision Trees-CART	0,660377	0,620253	0,671233	0,644737	0,7584652
453	Decision Trees-CART	0,748428	0,670886	0,791045	0,726027	0,8366297
454	Decision Trees-CART	0,716981	0,64557	0,75	0,693878	0,7957278
455	Decision Trees-CART	0,735849	0,683544	0,760563	0,72	0,8424051
456	Decision Trees-CART	0,81761	0,810127	0,820513	0,815287	0,8704905
457	Decision Trees-CART	0,779874	0,734177	0,805556	0,768212	0,8525316
458	Decision Trees-CART	0,767296	0,696203	0,808824	0,748299	0,8497627
459	Decision Trees-CART	0,735849	0,670886	0,768116	0,716216	0,8166139
460	Decision Trees-CART	0,779874	0,759494	0,789474	0,774194	0,8602848
461	Decision Trees-CART	0,691824	0,582278	0,741935	0,652482	0,7310127
462	Decision Trees-CART	0,767296	0,721519	0,791667	0,754967	0,8666139
463	Decision Trees-CART	0,698113	0,64557	0,71831	0,68	0,7719937
464	Decision Trees-CART	0,798742	0,746835	0,830986	0,786667	0,9166139
465	Decision Trees-CART	0,72327	0,658228	0,753623	0,702703	0,8059335

466	Decision Trees-CART	0,811321	0,797468	0,818182	0,807692	0,8940665
467	Decision Trees-CART	0,81761	0,78481	0,837838	0,810458	0,875712
468	Decision Trees-CART	0,754717	0,696203	0,785714	0,738255	0,8515032
469	Decision Trees-CART	0,754717	0,708861	0,777778	0,741722	0,859019
470	Decision Trees-CART	0,72327	0,6375	0,772727	0,69863	0,8077532
471	Decision Trees-CART	0,779874	0,7625	0,792208	0,77707	0,8328323
472	Decision Trees-CART	0,691824	0,675	0,701299	0,687898	0,7608386
473	Decision Trees-CART	0,72956	0,675	0,760563	0,715232	0,8102848
474	Decision Trees-CART	0,698113	0,65	0,722222	0,684211	0,7791139
475	Decision Trees-CART	0,779874	0,7625	0,792208	0,77707	0,8550633
476	Decision Trees-CART	0,710692	0,625	0,757576	0,684932	0,8273734
477	Decision Trees-CART	0,767296	0,6875	0,820896	0,748299	0,8439082
478	Decision Trees-CART	0,735849	0,725	0,74359	0,734177	0,8279272
479	Decision Trees-CART	0,761006	0,7125	0,791667	0,75	0,8456487
480	Decision Trees-CART	0,735849	0,725	0,74359	0,734177	0,8503165
481	Decision Trees-CART	0,72956	0,65	0,776119	0,707483	0,8091772
482	Decision Trees-CART	0,698113	0,6625	0,716216	0,688312	0,8141614
483	Decision Trees-CART	0,754717	0,7	0,788732	0,741722	0,8378956
484	Decision Trees-CART	0,761006	0,75	0,769231	0,759494	0,8523734
485	Decision Trees-CART	0,742138	0,6875	0,774648	0,728477	0,8240506
486	Decision Trees-CART	0,748428	0,7	0,777778	0,736842	0,8289557
487	Decision Trees-CART	0,748428	0,6625	0,80303	0,726027	0,840269
488	Decision Trees-CART	0,716981	0,6125	0,777778	0,685315	0,7981804
489	Decision Trees-CART	0,767296	0,6875	0,820896	0,748299	0,8678797
490	Decision Trees-CART	0,779874	0,7125	0,826087	0,765101	0,8606013
491	Decision Trees-CART	0,704403	0,675	0,72	0,696774	0,7956487
492	Decision Trees-CART	0,710692	0,675	0,72973	0,701299	0,7628956
493	Decision Trees-CART	0,792453	0,6875	0,873016	0,769231	0,8756329
494	Decision Trees-CART	0,754717	0,7125	0,780822	0,745098	0,8302215

495	Decision Trees-CART	0,798742	0,75	0,833333	0,789474	0,8814873
496	Decision Trees-CART	0,792453	0,7875	0,797468	0,792453	0,8454114
497	Decision Trees-CART	0,811321	0,7875	0,828947	0,807692	0,8821994
498	Decision Trees-CART	0,710692	0,75	0,697674	0,722892	0,7945411
499	Decision Trees-CART	0,72956	0,675	0,760563	0,715232	0,8209652

Anexo 3. Código fuente para la generación del conjunto de datos de mediciones de métricas por modelo

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
models2 = [
{
    'label': 'Logistic Regression',
    'model': LogisticRegression(),
},
{
    'label': 'Gradient Boosting',
    'model': GradientBoostingClassifier(n_estimators=100, max_depth=6,␣
 ↪random_state=42),
},
{
    'label':'Random Forest',
    'model': RandomForestClassifier(max_depth=6,␣
 ↪n_estimators=100,random_state=42),
},
{
    'label':'XGBoost',
    'model': XGBClassifier(objective='binary:logistic',max_depth=6,␣
 ↪n_estimators=100,random_state=42,verbosity=0)
},
{
    'label':'Decision Trees-CART',
    'model': DecisionTreeClassifier(random_state=42,␣
 ↪max_depth=6,criterion="gini"),
},
]
metricas_tr=[]
for m in models2:
    model = m['model']
    model.fit(X_sm, Y_sm)
    acc_train=cross_val_score(model,X_sm,Y_sm,cv=100,scoring="accuracy")
    sen_train=cross_val_score(model,X_sm,Y_sm,cv=100,scoring="recall")
    vpp_train=cross_val_score(model,X_sm,Y_sm,cv=100,scoring="precision")
    f1_train=cross_val_score(model,X_sm,Y_sm,cv=100,scoring="f1")
    roc_auc_train=cross_val_score(model,X_sm,Y_sm,cv=100,scoring="roc_auc")
    for a in zip(acc_train,sen_train,vpp_train,f1_train,roc_auc_train):
        metricas_tr.append([m['label'],a[0],a[1],a[2],a[3],a[4]])
df_metrics_tr_test_hipotesis = pd.DataFrame(metricas_tr, columns =␣
 ↪['Model','Accuracy','Sensitivity','PPV',"F1 Score","AUC"])
```

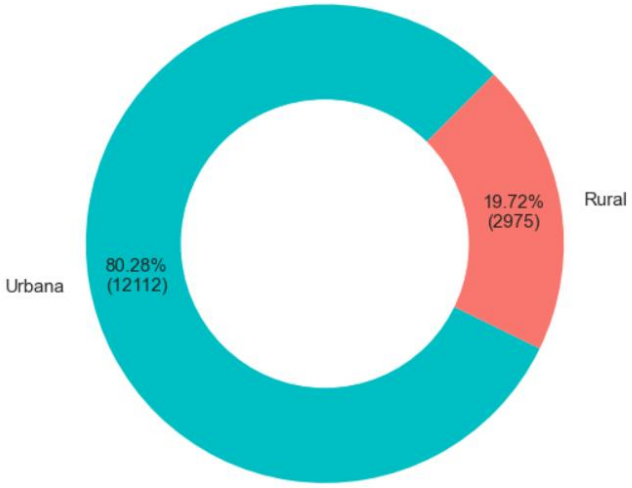
Anexo 4. Funciones auxiliares para el análisis exploratorio de datos

```
# Función que dibuja gráfico de donas
# Recibe la columna de un dataframe
def dia_donas(col):
    cantidad=col.value_counts()
    plt.style.use('seaborn-darkgrid')
    plt.figure(figsize=(9,9))
    clases=cantidad.index.tolist()
    area = cantidad.values.tolist()
    total = np.sum(area)

    def val_per(x):
        return '{:.2f}%\n({:.0f})'.format(x, total*x/100)
    labels = clases
    colors = ['#00BFC4','#F8766D','#8BC540']

    plt.pie(area , labels= labels , colors= colors , startangle=45 , autopct=val_per, pctdistance=0.80,textprops={'fontsize': 15})
    my_circle=plt.Circle( (0,0), 0.6, color='white')
    p=plt.gcf()
    p.gca().add_artist(my_circle)
    plt.show()
```

```
#Área al que pertenece
dia_donas(df_endo['P04'])
```




```
# Función que realiza diagrama de barras
def dia_barras(col,xlab="VARIABLE EN ESTUDIO",ylab="COUNT"):
    plt.figure(figsize=(20,7))
    ax=sns.countplot(x=col)
    cantidad=col.value_counts()
    total=np.sum(cantidad.values)

    plt.xticks(rotation = 90)
    plt.yticks(fontsize= 13)
    plt.xticks(fontsize= 12)
    plt.xlabel(col.name + "  " +str(xlab) , fontsize = 18)
    plt.ylabel(ylab , fontsize = 18)
    for p in ax.patches:
        ax.annotate('{0:.2%}'.format(p.get_height()/total),
                    (p.get_x() + p.get_width() / 2., p.get_height()),
                    ha = 'center', va = 'center',
                    xytext = (0, 9),
                    textcoords = 'offset points',
                    size=14)
    plt.show()
```

```
# Pais de procedencia del docente encuestado
dia_barras(df_endo['P101_1D'],'PAIS DE PROCEDENCIA','CANTIDAD DE ENCUESTADOS')
```

Anexo 5. Código fuente y conjunto de datos del proceso de modelado

Finalmente, pongo a disposición de la comunidad académica el repositorio donde se encuentra el código fuente y el conjunto de datos para el proceso de modelamiento de la satisfacción laboral docente.

URL del código fuente:
https://github.com/luisholgado/teacher_job_satisfaction/blob/main/teacher_job_satisfaction_code.ipynb

URL del conjunto de datos preprocesados:
https://data.mendeley.com/datasets/b7wbthz6hs/2

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo LUIS ALBERTO HOLGADO APAZA,
identificado con DNI 44076704 en mi condición de egresado de:

☐ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☒ **Programa de Doctorado**

EN CIENCIAS DE LA COMPUTACIÓN,

informo que he elaborado el/la ☐ **Tesis** o ☐ **Trabajo de Investigación** denominada:

"MODELAMIENTO DE LA SATISFACCIÓN LABORAL DE DOCENTES DE EDUCACIÓN BÁSICA MEDIANTE TÉCNICAS MACHINE LEARNING"

para la obtención de ☒**Grado,** ☐ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 04 de Julio del 2023

FIRMA (obligatoria)

Huella

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo LUIS ALBERTO HOLGADO APAZA,
identificado con DNI 44076704 en mi condición de egresado de:

☐ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☒ **Programa de Doctorado**
EN CIENCIAS DE LA COMPUTACIÓN,
informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

"MODELAMIENTO DE LA SATISFACCIÓN LABORAL DE DOCENTES DE EDUCACIÓN BÁSICA MEDIANTE TÉCNICAS MACHINE LEARNING"

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 04 de Julio del 2023

FIRMA (obligatoria)

Huella