

**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**

**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**

**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y  
SISTEMAS**



**“EVALUACIÓN COMPARATIVA EN EL DESEMPEÑO DE  
MODELOS DE MACHINE LEARNING EN LA PREDICCIÓN  
DE PAGADORES DEL IMPUESTO PREDIAL DE LA CIUDAD  
DE TINGO MARÍA PARA EL AÑO 2019”**

**Tesis**

**Para optar el título profesional de:**

**INGENIERO EN INFORMÁTICA Y SISTEMAS**

**PRESENTADO POR:**

**ANTONIO GABRIEL JUNCO ARANCIAGA**

**TINGO MARÍA – PERÚ**

**2023**



### PARTE 1. FASE INICIAL

Siendo las 11:20 horas del día **24 de octubre de 2023**; en la Sala de Conferencias de la FIIS, se instala el jurado calificador conformado por:

**Jurado 1:** Ing. Edwin Jesús Vega Ventocilla (presidente)

**Jurado 2:** Mg. Ronald Eduardo Ibarra Zapata

**Jurado 3:** Mg. Brian César Pando Soto

Oficializado mediante **RESOLUCIÓN N° 04-2023-D-FIIS-UNAS** del 03 de enero de 2023, para el proceso de sustentación del informe final de Tesis del bachiller **Antonio Gabriel JUNCO ARANCIAGA**, titulado: **"EVALUACIÓN COMPARATIVA EN EL DESEMPEÑO DE MODELOS DE MACHINE LEARNING EN LA PREDICCIÓN DE PAGADORES DEL IMPUESTO PREDIAL DE LA CIUDAD DE TINGO MARÍA PARA EL AÑO 2019"**. ASESOR: **Ing. Hubel Solis Bonifacio**. Se manifiesta que el bachiller cumple con los requisitos exigidos de Ley y se le invita a disertar su Tesis por espacio de 30 minutos, asimismo se dispondrá de igual tiempo para la absolver preguntas y sugerencias.

### PARTE 2. FASE DE PREGUNTAS Y RESULTADO

Culminada la exposición se inicia la fase de preguntas por parte del jurado calificador; también se invita a los asistentes a formular preguntas sobre el tema de Tesis.

Absueltas todas las peticiones, el jurado calificador procede a deliberar en privado la calificación y resultado.

Concluida la deliberación y en presencia del público, el jurado calificador anuncia que el resultado de la Sustentación de Tesis es: APROBADO POR UNANIMIDAD.

(NOTA: consignar una de la siguientes: DESAPROBADO, APROBADO POR MAYORIA o APROBADO POR UNANIMIDAD)

Con calificativo de: BUENO

(NOTA: consignar una de la siguientes: EXCELENTE, MUY BUENO, BUENO, DEFICIENTE, MUY DEFICIENTE)

Por lo que se comunicará a las instancias correspondientes para el trámite respectivo.

### PARTE 3. CONFORMIDAD

De todo lo mencionado se firma al pie en señal de conformidad, siendo las 12:50 horas se da por finalizada la ceremonia de Sustentación de Tesis.

Firma: 	Firma: 	Firma: 
<b>Jurado 1:</b> Edwin Jesús Vega Ventocilla	<b>Jurado 2:</b> Ronald Eduardo Ibarra Zapata	<b>Jurado 3:</b> Brian César Pando Soto
Firma: 	Firma: 	
<b>Sustentante:</b> Antonio Gabriel JUNCO ARANCIAGA	<b>Asesor:</b> Hubel Solis Bonifacio	



“Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junín y Ayacucho”

## CERTIFICADO DE SIMILITUD T.I. N° 008 - 2024 - CS-RIDUNAS

El Director de la Dirección de Gestión de Investigación de la Universidad Nacional Agraria de la Selva, quien suscribe,

### CERTIFICA QUE:

El Trabajo de Investigación; aprobó el proceso de revisión a través del software TURNITIN, evidenciándose en el informe de originalidad un índice de similitud no mayor del 25% (Art. 3° - Resolución N° 466-2019-CU-R-UNAS).

Programa de Estudio:

Ingeniería en Informática y Sistemas

Tipo de documento:

Tesis	<input checked="" type="checkbox"/>	Trabajo de Suficiencia Profesional	<input type="checkbox"/>
-------	-------------------------------------	------------------------------------	--------------------------

TÍTULO	AUTOR	PORCENTAJE DE SIMILITUD
EVALUACIÓN COMPARATIVA EN EL DESEMPEÑO DE MODELOS DE MACHINE LEARNING EN LA PREDICCIÓN DE PAGADORES DEL IMPUESTO PREDIAL DE LA CIUDAD DE TINGO MARÍA PARA EL AÑO 2019	ANTONIO GABRIEL JUNCO ARANCIAGA	<b>22 %</b> <b>Veintidós</b>

Tingo María, 09 de enero de 2024



UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA  
DIRECCIÓN DE GESTIÓN DE LA INVESTIGACIÓN

*Dr. Tomas Menacho Malqui*  
DIRECTOR

**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**  
**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**  
**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y**  
**SISTEMAS**



**“EVALUACIÓN COMPARATIVA EN EL DESEMPEÑO DE MODELOS DE  
MACHINE LEARNING EN LA PREDICCIÓN DE PAGADORES DEL IMPUESTO  
PREDIAL DE LA CIUDAD DE TINGO MARÍA PARA EL AÑO 2019”**

<b>Autor</b>	: Bach. Antonio Gabriel, Junco Aranciaga
<b>Asesor</b>	: Ing. Hubel Solis, Bonifacio
<b>Área de investigación</b>	: Ingeniería de software y computación.
<b>Línea de investigación</b>	: Computación Aplicada
<b>Eje temático</b>	: Machine Learning.
<b>Lugar de ejecución</b>	: Tingo María, Huánuco
<b>Duración</b>	: 6 meses
<b>Financiamiento</b>	: S/ 1,500.00

**Tingo María – Perú. 2023**

VICERRECTORADO DE INVESTIGACION  
OFICINA DE INVESTIGACION



UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

REGISTRO DE TESIS PARA LA OBTENCION DEL  
TITULO UNIVERSITARIO, INVESTIGACIÓN DOCENTE  
Y TESISISTA

(Resol. N° 113-2019-CU-R-UNAS)

I. Datos Generales de Pregrado

<b>Universidad</b>	:	Universidad Nacional Agraria de la Selva.
<b>Facultad</b>	:	Ingeniería en Informática y Sistemas.
<b>Título de tesis</b>	:	Evaluación comparativa en el desempeño de modelos de machine learning en la predicción de pagadores del impuesto predial de la ciudad de tingo maría para el año 2019.
<b>Autor</b>	:	Antonio Gabriel, Junco Aranciaga.
<b>Asesor de tesis</b>	:	Solis Bonifacio, Hubel
<b>Escuela Profesional</b>	:	Informática y Sistemas.
<b>Área de investigación</b>	:	Ingeniería de Software y computación.
<b>Línea(s) de investigación</b>	:	Computación Aplicada.
<b>Eje Temático</b>	:	Machine Learning.
<b>Lugar de ejecución</b>	:	Tingo María, Huánuco.
<b>Duración</b>	:	Inicio : Abril 2023 Término : Octubre 2023
<b>Financiamiento</b>	:	FEDU : S/0.00 Propio : S/1,500.00 Otros : S/.0.00

Tingo María, Perú, Enero 2024.

A handwritten signature in blue ink, appearing to be 'AG'.

Antonio Gabriel, Junco Aranciaga

Tesista

A handwritten signature in blue ink, appearing to be 'HB'.

Hubel, Solis Bonifacio

Asesor

## **DEDICATORIA**

A Dios en primer lugar por permitirme la vida brindándome salud y sabiduría para lograr llegar hasta este momento muy importante de mi vida y formación profesional.

A mi papá Manuel, por su apoyo incondicional en mi formación profesional ya que sin la inspiración, entusiasmo y ayuda que me diste, nunca me hubiera convertido en la persona que soy. Gracias, papá.

A mi querida Mamá Luisa, que me crió como hijo cuya dedicación y amor han sido la base de mi vida y éxito académico. Sin tu apoyo incondicional y tus sabias palabras, esta tesis no habría sido posible. Gracias por ser mi guía, mi refugio y mi inspiración en cada paso del camino. Deseo que estas páginas sean un tributo a tu amor y sacrificio, y espero hacerte sentir orgullosa con este logro.

A mi mamá Isabel, mi amor Gianelly, mis hijos Ayxa y Daryl por ser mi inspiración y motivo de seguir adelante.

## **AGRADECIMIENTO**

Agradezco a Dios por bendecirme en la vida, por guiarme a lo largo de mi existencia y brindarme una linda familia.

A mi pareja Gianelly Aguirre Callan, por entenderme en todo, gracias a ella porque en todo momento fue un apoyo incondicional en mi vida, fue la felicidad encajada en una sola persona, fue mi todo reflejado en otra persona a la cual yo amo demasiado, y por la cual estoy dispuesto a enfrentar todo y en todo momento.

A los docentes de la Facultad de Ingeniería en Informática y Sistemas y jurado calificador, por los estímulos y recomendaciones para elaborar y culminar la presente investigación.

A mi tío José que desde muy niño me inculcó al estudio y fue sin duda el maestro en mis enseñanzas y motivación en el desarrollo de mi formación académica.

A mis abuelitas Magdalena y Fortunata aunque sus partidas me han dejado un gran dolor, su amor y su guía siempre me acompañarán

A mi tío Augusto que fue un tío maravilloso y siempre te recordaremos con cariño.

A mi familia por su amor incondicional, paciencia, comprensión y apoyo emocional durante todo este tiempo, así como por ser mi mayor fuente de inspiración y motivación.

## TABLA DE CONTENIDO

I.	INTRODUCCIÓN.....	1
1.1.	Objetivo general .....	3
1.2.	Objetivos específicos.....	3
II.	REVISIÓN DE LITERATURA .....	7
2.1.	Marco teórico.....	7
2.2.	Estado del arte .....	21
III.	METODOLOGÍA.....	25
3.1.	Tipo y nivel de investigación.....	25
3.2.	Población y muestra .....	25
3.3.	Hipótesis .....	26
3.4.	Variables de investigación.....	26
3.5.	Operacionalización de variables.....	27
3.6.	Herramientas de recolección de datos .....	28
3.7.	Metodología.....	28
IV.	RESULTADOS Y DISCUSIÓN .....	32
4.1.	Descripción e influencia de características.....	32
4.2.	Evaluación de la Exactitud .....	40
4.3.	Evaluación de la Precisión.....	44
4.4.	Evaluación del Recall .....	49
4.5.	Puntuación F1 o Desempeño .....	52
V.	CONCLUSIÓN.....	55
VI.	REFERENCIAS BIBLIOGRÁFICAS .....	58
	ANEXOS .....	63



## ÍNDICE DE FIGURAS

Figura	Pag
1. Esquema de las disciplinas que componen a la ciencia de datos .....	12
2. Arquitectura básica de una red neuronal .....	18
3. Ingreso económico de los diferentes tipos de contribuyentes.....	32
4. Situación económica como causa de impago/impuntualidad/pago de tributos.....	33
5. Percepción de los contribuyentes del cumplimiento de las autoridades para solucionar problemas de la ciudad.....	34
6. Percepción de los contribuyentes con respecto a la corrupción de los funcionarios.	35
7. Percepción de los contribuyentes con respecto a la utilidad de los tributos. ....	36
8. Conocimiento de los contribuyentes con respecto a las sanciones tributarias.....	37
9. Priorización de los contribuyentes con respecto a los gastos necesarios del hogar.	38
10. Estabilidad laboral de los contribuyentes. ....	39
11. Percepción de los contribuyentes con respecto a la utilidad de los tributos. ....	40
12. Matriz de confusión, Precisión y Recall de los modelos: a) Decisión Tree, b) Support Vector Machine, c) K - Nearest Neighbors, d) Naive, e) Ensemble y f) Redes Neuronales. ....	44

## INDICE DE TABLAS

Tabla	Pag
1. Cálculo del impuesto predial .....	9
2. Variables socioeconómicas y de percepción que influyen en el tipo de pagador del contribuyente.....	40
3. ANOVA de Friedman de la Exactitud de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de $\alpha = 0.05$ . ....	43
4. ANOVA de Friedman de la Precisión para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de $\alpha = 0.05$ . ....	48
5. ANOVA de Friedman de la Sensibilidad (Recall) para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de $\alpha = 0.05$ . ....	51
6. ANOVA de Friedman de la Puntuación F1 o Desempeño para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de $\alpha = 0.05$ . ....	53

## RESUMEN

La evasión fiscal y el retraso en el pago de impuestos son problemas persistentes a nivel mundial que afectan la recaudación de ingresos y la provisión de servicios sociales. En contextos locales, como en la ciudad de Tingo María, estos problemas se manifiestan de manera más aguda. Según los datos de la Municipalidad Provincial de Leoncio Prado, de los 12,821 contribuyentes registrados, 8,269 no pagan sus impuestos y 187 pagan con retraso, lo que representa una pérdida significativa de ingresos para la municipalidad. Para abordar esta problemática, se implementó una metodología de clasificación de pagadores basada en la recopilación de información socioeconómica y de percepción a través de encuestas telefónicas. Se seleccionó una muestra de 372 contribuyentes, distribuidos equitativamente entre las categorías de pagadores puntuales, impuntuales y no pagadores. Los datos recopilados se sometieron a un análisis descriptivo utilizando un mapa de calor y la prueba de Chi<sup>2</sup>, revelando una correlación significativa entre el ingreso económico y la puntualidad en el pago de impuestos. Para mejorar la identificación temprana de contribuyentes impuntuales o no pagadores, se evaluaron seis modelos de Machine Learning en términos de exactitud, precisión y sensibilidad. Los modelos de K-Nearest Neighbors (KNN) y Redes Neuronales (RNN) demostraron ser los más precisos, con una exactitud superior al 97%. Estos hallazgos sugieren que las autoridades fiscales podrían utilizar estos modelos para tomar medidas proactivas y prevenir la evasión fiscal. Además, se recomienda la inclusión de más variables socioeconómicas y de percepción para mejorar aún más la precisión de estos modelos. La implementación de estas recomendaciones podría permitir a las autoridades fiscales optimizar sus recursos, mejorar la recaudación de impuestos y, en última instancia, mejorar la provisión de servicios sociales en la ciudad de Tingo María.

**Palabras clave:** evasión fiscal, machine learning, recaudación de impuestos, métricas de rendimiento, percepción socioeconómica

## ABSTRACT

For this research, socioeconomic information and the perception of 372 tax payers in Tingo Maria, [Peru], was collected through phone surveys. A descriptive analysis was done and the influence of these factors on the classification of the type of tax payer: non-payer, late, and punctual, was evaluated using the chi-squared test. The data that was collected was used to train the machine learning models, which included: decision tree, SVM, KNN, Naive Bayes, ensemble, and RNN; and the metrics of exactness, precision, sensibility, and F-score were measured. A ten-fold k-fold cross-validation, and the hyperparameters were optimized. The differences in the yield of the models were determined using Friedman's ANOVA and Nemenyi's post-hoc tests. For the results, the socioeconomic factors and perceptions notably influenced the classification of the tax payers ( $p$ -value  $< 0.01$ ). For the prediction, the KNN and RNN models stood out with an exactness of  $97.04\% \pm 0.30$  and  $96.89\% \pm 0.14$ . For the precision, RNN led in "non-payers" at  $93.85 \pm 0.25$ , and KNN in "late" at  $99.10 \pm 0.12$ ; meanwhile, Naive Bayes stood out for "punctual" at  $99.90 \pm 0.04$ . For the sensibility, RNN had  $95.55 \pm 0.31$  for "non-payers" and  $99.92 \pm 0.04$  for "punctual"; KNN achieved  $95.67 \pm 0.05$  for "late." For the f-score, KNN and RNN had the best results for "non-payers" at  $99.13 \pm 0.15$  and for "punctual" at  $99.73 \pm 0.15$ , respectively. Using Friedman's ANOVA, it was concluded that significant differences existed between the algorithms for the evaluation of the tax payers ( $p$ -value  $< 0.01$ ).

Keywords: machine learning, property tax, yield metrics, Friedman, socioeconomic factors

## I. INTRODUCCIÓN

El impuesto predial es una fuente de ingresos importante para las ciudades, ya que se utiliza para financiar los servicios públicos y obras que benefician a toda la comunidad. Cuando los ciudadanos no pagan el impuesto predial, se reducen los recursos disponibles para la ciudad, lo que puede tener consecuencias negativas en términos de la calidad de vida de la comunidad. De acuerdo con información de la OCDE, la recaudación por Impuesto Predial en el Perú equivale al 0,2% del PBI, la mitad del promedio de América Latina (0,4%) y menos de un quinto de los países OCDE (1,1%) (BCRP, 2021; OCDE, 2020). La Municipalidad de Lima informó que la recaudación por concepto de impuesto predial y arbitrios en el 2020 fue de S/ 1,800 millones, lo que representa una disminución del 13% respecto al 2019. La baja recaudación de impuestos prediales afecta directamente el financiamiento de servicios públicos como la limpieza pública, seguridad ciudadana, parques y jardines, entre otros (Cayatopa Rivera, 2021).

Otro problema que se puede generar con el impago del impuesto predial es que si los ciudadanos perciben que el gobierno no puede recaudar los impuestos adecuadamente, puede generar una pérdida de confianza en el gobierno y en sus capacidades para mantener y mejorar los servicios públicos. Si solo algunos ciudadanos pagan sus impuestos prediales, mientras que otros no lo hacen, puede generar una sensación de desigualdad social. Aquellos que pagan sus impuestos pueden sentirse perjudicados por tener que financiar los servicios públicos y la infraestructura mientras que otros no contribuyen. Esto podría generar un efecto en cadena en la voluntad ciudadana para cumplir con sus pagos de sus impuestos prediales (Smolka & De Cesare, 2012).

Los contribuyentes morosos también representan un problema porque el retraso en el pago del impuesto predial puede causar problemas a la administración municipal, ya que puede retrasar el proceso de planificación y ejecución de proyectos y servicios.

En Perú el impuesto predial representa un 72% de los ingresos tributarios municipales, por ello se debe crear una “cultura tributaria” pertinente para su exitosa recaudación (Romero Carazas et al., 2021).

Una de las maneras de gestionar la recaudación de los tributos es conociendo de manera anticipada quienes serán los contribuyentes que pagarán puntualmente, los que pagarán de manera impuntual y los que no pagarán. Con esta información se podría

ahorrar costos en la gestión y recaudación de impuestos, ya que se pueden enfocar los recursos en aquellos contribuyentes que tienen más probabilidades de no cumplir con sus obligaciones tributarias, de esta manera los esfuerzos se canalizarían para este tipo de contribuyentes, por otra parte, este conocimiento anticipado de los pagadores, permite conocer también de manera anticipada cual será el monto recaudado aproximado y de esa manera el municipio puede planificar mejor su presupuesto y asignar recursos de manera más efectiva (FMI, 2023)

Existen diversas herramientas para predecir el tipo de pagador tributario, hoy entre las principales podemos mencionar a los modelos estadísticos y a la minería de datos y de machine learning o aprendizaje automático. Una de las principales ventajas del machine learning sobre otras herramientas es su capacidad para analizar grandes conjuntos de datos y detectar patrones y relaciones complejas. En el caso de la predicción de tipos de pagadores de impuesto predial, los modelos de machine learning pueden analizar una gran cantidad de datos socioeconómicos de los contribuyentes, identificar las características más importantes para la clasificación de pagadores puntuales, impuntuales y no pagadores, y utilizar estas características para predecir con alta precisión el tipo de pagador de cada contribuyente. Además, los modelos de machine learning pueden ser entrenados para mejorar su precisión a medida que se recopilan más datos, lo que permite una mejora continua de las predicciones. También pueden manejar diferentes tipos de datos, incluyendo datos numéricos y categóricos, y pueden ser implementados en diferentes plataformas y lenguajes de programación.

A continuación, se menciona de manera breve el contenido por capítulo. Introducción: Este capítulo establece el marco de la tesis, presentando el problema de la evasión fiscal y la necesidad de un sistema de predicción para identificar a los contribuyentes potencialmente evasores. Se establecen los objetivos generales y específicos, así como la hipótesis de que los modelos de aprendizaje automático pueden predecir con precisión la evasión fiscal. Revisión de literatura: Este capítulo presenta una revisión exhaustiva de la literatura existente sobre la evasión fiscal y los modelos de aprendizaje automático. Se discuten varios estudios internacionales que han utilizado el aprendizaje automático para predecir la evasión fiscal, destacando sus hallazgos y limitaciones. Metodología: Este capítulo describe la metodología utilizada para desarrollar y evaluar los modelos de aprendizaje automático. Se discute el proceso de

recopilación de datos, la selección de características, la construcción del modelo y la evaluación del rendimiento del modelo. Resultados y discusión: Este capítulo presenta los resultados obtenidos de los modelos de aprendizaje automático. Se discuten los factores que influyen en la evasión fiscal, como el ingreso económico, y se evalúa la precisión, la exactitud y el recall de los modelos. Se muestra que el nivel de ingreso económico de los contribuyentes está relacionado con su capacidad para realizar pagos puntuales. Conclusiones: Este capítulo resume los hallazgos de la tesis, destacando la eficacia de los modelos de aprendizaje automático para predecir la evasión fiscal. Se discuten las implicaciones de estos hallazgos para las políticas fiscales y se sugieren áreas para futuras investigaciones.

Existen una gran variedad de modelos de machine learning y cada modelo puede tener mejor desempeño para cada problema en particular. Por lo expuesto anteriormente se plantea la siguiente interrogante ¿Existe diferencia en el desempeño de modelos en la predicción de pagadores del impuesto predial de la ciudad de Tingo María para el año 2019?

### **1.1. Objetivo general**

Evaluar la diferencia en el desempeño de modelos de Machine Learning en la predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

### **1.2. Objetivos específicos**

Evaluar la diferencia en la exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

Evaluar la diferencia en la precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

Evaluar la diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

## **Importancia**

La predicción de los pagadores del impuesto predial es un componente crucial en la administración de los ingresos municipales y en el diseño de políticas públicas eficaces. Los ingresos procedentes del impuesto predial son fundamentales para los gobiernos locales, ya que financian servicios públicos esenciales como la educación, la sanidad, la infraestructura y la seguridad. Por lo tanto, comprender qué factores socioeconómicos pueden influir en el pago del impuesto predial es esencial para garantizar una gestión eficaz y sostenible de los recursos públicos.

En primer lugar, este estudio ofrece un método sistemático para identificar los factores socioeconómicos que más probablemente determinen el pago del impuesto predial. Al hacerlo, proporciona a los responsables de la toma de decisiones información valiosa que puede ayudarles a identificar a los grupos de población que pueden necesitar apoyo adicional o a los que podrían dirigirse estrategias de cumplimiento más enfocadas.

En segundo lugar, el uso de varios modelos de aprendizaje automático para predecir el comportamiento de pago puede ayudar a mejorar la eficiencia de las operaciones municipales. Al anticipar quiénes son los pagadores puntuales, los no pagadores y los pagadores atrasados, los gobiernos pueden optimizar su gestión de recursos y diseñar intervenciones más efectivas para mejorar el cumplimiento del pago de impuestos.

Además, este estudio puede arrojar luz sobre las desigualdades socioeconómicas y cómo estas pueden afectar al cumplimiento del pago de impuestos. Con ello, puede contribuir a la creación de políticas públicas más equitativas y a la reducción de las brechas socioeconómicas.

Finalmente, los hallazgos de este estudio podrían tener aplicaciones más allá de la administración del impuesto predial. Los factores socioeconómicos y los modelos de predicción podrían aplicarse también a otras formas de tributación o a la prestación de servicios públicos. En general, este estudio puede hacer una contribución significativa a la forma en que se administran y se recaudan los impuestos, lo que, en última instancia, puede conducir a una mejora de los servicios públicos y a una mayor equidad en la sociedad.



### **Justificación teórica**

La recaudación del impuesto predial es una fuente importante de ingresos para las municipalidades, y su gestión efectiva depende en gran medida de la capacidad de las autoridades para identificar y clasificar a los contribuyentes según su capacidad y disposición para pagar el impuesto. En este contexto, los modelos de machine learning se han convertido en una herramienta valiosa para la predicción de pagadores del impuesto predial, ya que pueden analizar grandes conjuntos de datos socioeconómicos y detectar patrones y relaciones complejas que pueden ser difíciles de detectar mediante métodos tradicionales.

Sin embargo, hay una gran variedad de modelos de machine learning que se pueden aplicar para la predicción de pagadores del impuesto predial, y cada modelo tiene sus propias fortalezas y debilidades. Por lo tanto, es importante evaluar y comparar el desempeño de diferentes modelos de machine learning en términos de su capacidad para predecir con precisión los pagadores puntuales, impuntuales y no pagadores.

### **Justificación práctica (tecnológica)**

La justificación tecnológica de esta investigación radica en el aprovechamiento de las técnicas de aprendizaje automático para enfrentar el desafío creciente de la evasión del impuesto predial. Con la digitalización y la creciente recopilación de datos fiscales y socioeconómicos, el volumen y la complejidad de estos datos superan las capacidades de los métodos tradicionales de análisis y predicción.

Los algoritmos de aprendizaje automático son capaces de manejar grandes conjuntos de datos y detectar patrones sutiles y relaciones complejas que pueden ser difíciles de identificar mediante técnicas de análisis convencionales. Además, la capacidad de estos algoritmos para aprender de los datos y mejorar con el tiempo, ofrece la posibilidad de crear sistemas de predicción más precisos y eficientes.

Desde una perspectiva práctica, la aplicación de estas técnicas puede llevar a una gestión más eficaz de los recursos públicos y a una mayor capacidad de las autoridades para anticipar y prevenir la evasión fiscal. Esto puede resultar en una recaudación de impuestos más equitativa y eficiente, lo que a su vez puede tener un impacto significativo en la calidad de los servicios públicos y la equidad socioeconómica.

Además, esta investigación puede sentar las bases para el desarrollo de nuevas herramientas y técnicas de análisis fiscal que incorporen el aprendizaje automático, lo que podría abrir nuevas vías para la innovación en la administración fiscal.

### **Alcance**

El alcance de esta investigación se centra en la evaluación comparativa del desempeño de modelos de machine learning en la predicción de los pagadores del impuesto predial específicamente en la ciudad de Tingo María para el año 2019. El estudio se dirige a investigadores, académicos y profesionales interesados en el ámbito de la administración tributaria y la predicción del comportamiento de pago de impuestos.

Los resultados de esta investigación pueden ser utilizados por las autoridades fiscales de la ciudad de Tingo María para mejorar la gestión y la eficiencia en la recaudación del impuesto predial. También puede ser relevante para profesionales y expertos en el campo de la calificación crediticia y la evaluación de riesgos, ya que la predicción del comportamiento de pago de impuestos puede estar relacionada con la capacidad de los contribuyentes para cumplir con sus obligaciones financieras.

Además, el alcance de este estudio puede extenderse a otras ciudades o regiones que enfrenten desafíos similares en la recaudación del impuesto predial. Los resultados y las metodologías utilizadas pueden servir como base para futuras investigaciones y aplicaciones en el campo de la administración tributaria y la aplicación de modelos de machine learning en la predicción de pagadores de impuestos.

## **II. REVISIÓN DE LITERATURA**

### **2.1. Marco teórico**

A continuación, se presenta el marco teórico de la investigación, sobre los fundamentos de la recaudación municipal y como puede gestionarse y apoyarse a partir de la inteligencia artificial.

#### **2.1.1. Recaudación Municipal Tributaria Predial**

La recaudación municipal se refiere al proceso mediante el cual un gobierno local, como un municipio o una ciudad, obtiene ingresos para financiar sus operaciones y proyectos. Estos ingresos son recaudados principalmente a través de impuestos y otras fuentes de ingresos locales.

Los municipios recaudan impuestos de varias formas, que pueden incluir impuestos a la propiedad, impuestos sobre las ventas, impuestos sobre vehículos, impuestos sobre la renta o ingresos, y otros impuestos o tasas locales específicas. Estos impuestos se utilizan para financiar una variedad de servicios y programas municipales, como el mantenimiento de carreteras y calles, servicios de agua y alcantarillado, servicios de recolección de basura, parques y espacios públicos, servicios de policía y bomberos, y programas sociales, entre otros.

##### **2.1.1.1. ¿Qué es el impuesto predial?**

El impuesto predial es un tributo que grava el valor de todos los predios urbanos y rústicos en base a su autoevaluación, el cual se obtiene aplicando los aranceles y precios unitarios de construcción aprobados por el Ministerio de Vivienda, Construcción y Saneamiento (MVCS) todos los años. La recaudación, administración y fiscalización del impuesto predial está a cargo de la municipalidad distrital donde se ubica el predio (Meza, 2019)

De acuerdo con el artículo 8 del TUO de la ley de tributación municipal, el impuesto predial grava el valor de los predios urbanos y rústicos. Se considera previos a los terrenos incluyendo los ganados al mar los ríos y los otros espejos de agua y edificaciones e instalaciones fijas y permanentes que son parte de los terrenos que no pueden ser separados sin alterar deteriorar o destruir la edificación (MEF et al., 2016)

### **2.1.1.2. ¿Quiénes se encuentran obligados al pago del Impuesto Predial?**

Se encuentran obligados al pago del Impuesto Predial las personas naturales y jurídicas que sean propietarias de los predios gravados al 1° de enero de cada año. En caso de transferir el predio, el comprador asumirá la condición de contribuyente y estará obligado al pago del impuesto a partir del año siguiente de producida la transferencia. Los condóminos o copropietarios están en la obligación de comunicar la parte proporcional del predio que les corresponde (porcentaje de propiedad) a la municipalidad de su distrito. Sin embargo, la municipalidad puede exigir a cualquiera de ellos el pago total del impuesto. Sólo cuando la administración tributaria no pueda determinar quién es el propietario, se encuentran obligados al pago, en calidad de responsables solidarios, los poseedores o tenedores del predio, sin perjuicio de su derecho a reclamar el pago a los respectivos contribuyentes (MEF et al., 2016)

### **2.1.1.3. ¿Sobre qué base se calcula el Impuesto Predial?**

El impuesto se calcula sobre el valor total de los predios del contribuyente ubicados en cada jurisdicción distrital. A efectos de determinar el valor total de los predios, se aplicarán los valores arancelarios de terrenos y valores unitarios oficiales de edificación vigentes al 31 de octubre del año anterior y las tablas de depreciación por antigüedad y estado de conservación, que aprueba anualmente el Ministro de Vivienda, Construcción y Saneamiento mediante Resolución Ministerial. Las instalaciones fijas y permanentes serán valorizadas por el contribuyente de acuerdo con la metodología aprobada en el Reglamento Nacional de Tasaciones y utilizará una depreciación según antigüedad y estado de conservación. Dicha valorización está sujeta a fiscalización posterior por parte de la municipalidad respectiva. En el caso de terrenos que no hayan sido considerados en los planos básicos arancelarios oficiales, el valor de los mismos será estimado por la municipalidad Distrital respectiva o, en defecto de ella, por el contribuyente, tomando en cuenta el valor arancelario más próximo a un terreno de iguales características (MEF et al., 2016)

### **2.1.1.4. ¿Cómo se calcula el Impuesto Predial?**

El impuesto se calcula aplicando la siguiente escala progresiva acumulativa al valor total de los predios ubicados dentro del distrito:

**Tabla 1.** Cálculo del impuesto predial

Tramo de Autoevalúo	Alícuota
Hasta 15 UIT	0,2%
Más de 15 UIT y hasta 60 UIT	0,6%
Más de 60 UIT	1,0%

### 2.1.2. Factores socioeconómicos que influyen en el pago de impuesto predial

El pago de impuestos prediales es una obligación fiscal que tienen los propietarios de bienes raíces, y su cumplimiento es esencial para el mantenimiento de las finanzas públicas y el desarrollo del país. Sin embargo, el pago de impuestos prediales puede verse afectado por diversos factores socioeconómicos que influyen en el comportamiento de los contribuyentes. Por ejemplo, en la tesis de (Montero Flores, 2022), los factores que más influyen y determinan la probabilidad de cumplimiento de pago son el ahorro, condición laboral, retiro de AFP e ingreso promedio familiar. A continuación, una breve explicación sobre cada uno de estos factores:

**Ahorro:** El autor menciona que el ahorro es un factor importante en el cumplimiento del pago del impuesto predial ya que las personas que tienen un mayor nivel de ahorro tienen una mayor capacidad para hacer frente a sus obligaciones financieras.

**Condición laboral:** El autor indica que la condición laboral también influye en el cumplimiento del pago del impuesto predial. Las personas con empleos estables y bien remunerados tienen una mayor capacidad para pagar sus impuestos.

**Retiro de AFP:** El autor menciona que el retiro de AFP también puede influir en el cumplimiento del pago del impuesto predial. Si las personas retiran grandes sumas de dinero de sus cuentas AFP, pueden tener dificultades para hacer frente a sus obligaciones financieras.

**Ingreso promedio familiar:** El autor indica que el ingreso promedio familiar es un factor importante en el cumplimiento del pago del impuesto predial. Las familias con mayores ingresos tienen una mayor capacidad para pagar sus impuestos.

Por otra parte (Chávez Alvarez, 2021), concluyó que el grado de cultura tributaria y el nivel socioeconómico influyen significativamente en la recaudación de impuestos

prediales, es importante mencionar que el grado de cultura tributaria estuvo representado por el grado de difusión de la importancia de los tributos por parte de la municipalidad, y el grado de conocimiento de las personas sobre la utilidad de los tributos, por otra parte el nivel socioeconómico estuvo representado por el ingreso económico, si la persona provenía de una zona rural o urbana y del nivel educativo. Otro factor importante es la corrupción percibida, según (Alvarado Baena, 2023) la corrupción tiene un efecto negativo en la recaudación de impuestos. Las personas pagan sobornos para evitar el pago de impuestos y cuando los contribuyentes perciben que sus gobiernos son corruptos, son más propensos a evadir este pago. Si los ciudadanos confían en sus gobiernos y creen que sus impuestos se están utilizando de manera efectiva, es más probable que paguen sus impuestos y que inviertan en el país.

### **2.1.3. Fundamentos teóricos de Machine Learning**

Machine Learning se basa en una combinación de conceptos matemáticos, estadísticos y computacionales. El aprendizaje automático utiliza técnicas de probabilidad, estadística, teoría de la información, optimización y generalización para construir modelos que puedan aprender de los datos y mejorar su rendimiento en una tarea específica. La capacidad de explicar y comunicar los resultados obtenidos es también fundamental para la aplicación exitosa del aprendizaje automático en diferentes campos de la ciencia, la industria y la sociedad en general. La ciencia de datos y el Machine Learning están intrínsecamente relacionados, y el Machine Learning es una herramienta fundamental en la ciencia de datos para analizar, procesar y extraer información útil de grandes conjuntos de datos (Robert, 2014).

**Teoría de la probabilidad y estadística:** La teoría de la probabilidad y estadística es fundamental en Machine Learning, ya que proporciona los fundamentos matemáticos para el análisis de datos y la construcción de modelos de aprendizaje automático. La probabilidad y la estadística son herramientas esenciales para la inferencia, la estimación y la predicción basadas en datos. En Machine Learning, la teoría de la probabilidad se utiliza para modelar la incertidumbre y la variabilidad presentes en los datos (Robert, 2014).

**Teoría de la información:** La teoría de la información es una disciplina matemática que estudia la cuantificación, almacenamiento y transmisión de la información. En Machine

Learning, la teoría de la información se utiliza para medir la complejidad y la capacidad de un modelo de aprendizaje automático para transmitir información. La entropía, la información mutua y la divergencia de Kullback-Leibler son conceptos importantes en la teoría de la información utilizados en Machine Learning (Robert, 2014).

**Técnicas de optimización matemática:** Las técnicas de optimización matemática son fundamentales en Machine Learning, ya que se utilizan para ajustar los parámetros de los modelos de aprendizaje automático. La optimización de funciones de costo es una tarea común en el entrenamiento de modelos de Machine Learning. Las técnicas de optimización incluyen el descenso del gradiente, el método de Newton-Raphson, el método de Levenberg-Marquardt, entre otros (Robert, 2014).

**Noción de generalización:** La noción de generalización en Machine Learning se refiere a la capacidad de un modelo de aprender a partir de los datos de entrenamiento y aplicar ese conocimiento a nuevos datos no vistos previamente. La generalización es uno de los objetivos principales en Machine Learning y se logra a través de la selección de modelos apropiados, el control del sobreajuste y la utilización de técnicas de validación y prueba de modelos (Robert, 2014).

**Modelos matemáticos complejos:** Los modelos matemáticos complejos son aquellos que contienen múltiples capas, neuronas y parámetros, y son capaces de modelar relaciones no lineales entre las variables de entrada y salida. Estos modelos incluyen redes neuronales, árboles de decisión, SVM, entre otros. Los modelos matemáticos complejos son utilizados en tareas de Machine Learning como la clasificación de imágenes, el procesamiento de lenguaje natural y la detección de fraudes financieros (Robert, 2014).



**Figura 1.** Esquema de las disciplinas que componen a la ciencia de datos

#### **2.1.4. Áreas de aplicación de Machine Learning**

Machine Learning es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos matemáticos que pueden aprender de los datos y hacer predicciones o tomar decisiones. Desde su aparición en la década de 1950, el Machine Learning ha evolucionado y ha encontrado una gran cantidad de aplicaciones en diversos campos. A continuación, se describen algunos de los campos de aplicación más importantes de Machine Learning.

**Finanzas:** Machine Learning se utiliza en el sector financiero para analizar grandes cantidades de datos y hacer predicciones precisas. Esto incluye la predicción de riesgos crediticios, la detección de fraudes financieros, la optimización de carteras de inversión y el análisis de sentimientos del mercado (Goodell et al., 2021).

**Medicina:** La medicina es otro campo en el que Machine Learning ha encontrado una amplia gama de aplicaciones. Esto incluye la identificación de patrones de enfermedades, la predicción de la eficacia de los tratamientos, la identificación de genes relacionados con enfermedades y la detección de cáncer (Rajkomar et al., 2019).

**Automatización de procesos:** Machine Learning se utiliza para automatizar una amplia gama de procesos empresariales y operativos. Esto incluye la automatización de procesos



de fabricación, la optimización de la cadena de suministro y la predicción del mantenimiento (McKinsey, 2023).

**Marketing:** El Machine Learning se utiliza en el marketing para predecir el comportamiento del cliente, personalizar las ofertas y las campañas publicitarias, y mejorar la satisfacción del cliente (Verma et al., 2021).

**Agricultura:** El Machine Learning se utiliza en la agricultura para predecir el rendimiento de los cultivos, mejorar la calidad del suelo y optimizar los sistemas de riego (Liakos et al., 2018).

**Educación:** El Machine Learning se utiliza en la educación para personalizar el aprendizaje y mejorar los resultados de los estudiantes. Esto incluye la identificación de áreas de debilidad de los estudiantes y la sugerencia de actividades de aprendizaje específicas (Fahd et al., 2022).

### **2.1.5. Algoritmos de Machine Learning**

Los algoritmos de machine learning (aprendizaje automático) son una clase de algoritmos diseñados para permitir a las máquinas aprender y tomar decisiones o hacer predicciones basadas en datos sin ser explícitamente programadas. Estos algoritmos son fundamentales en el campo de la inteligencia artificial y se utilizan en una amplia variedad de aplicaciones.

En lugar de seguir instrucciones específicas, los algoritmos de machine learning aprenden de los datos y encuentran patrones, tendencias y relaciones en ellos. Estos algoritmos son capaces de aprender de forma autónoma y mejorar su rendimiento a medida que se les proporciona más información.

#### **2.1.5.1. Modelos de regresión**

Los modelos basados en la regresión se utilizan principalmente en la detección de fraudes en la contabilidad financiera. La mayoría de ellos se basan en la regresión logística, la regresión logística por pasos, el método de toma de decisiones multicriterio y la beta 2 generalizada exponencial (EGB2) (Ngai et al., 2011).

El modelo logístico es un modelo lineal generalizado que se utiliza para la regresión binomial en el que las variables predictoras pueden ser numéricas o categóricas (Spathis, 2002; Yeh & hui Lien, 2009). Se utiliza principalmente para resolver los problemas causados por los seguros y el fraude corporativo. Algunas de las investigaciones han sugerido un modelo basado en la regresión logística para predecir la presencia de fraude en los estados financieros. El método estadístico de regresión logística puede detectar estados financieros falsificados de manera eficiente (Spathis, 2002)

En (Bell & Carcello, 2000), Bell y Carcello propusieron un modelo de regresión logística para estimar la probabilidad de que un cliente de auditoría presentara informes financieros fraudulentos. El modelo estaba condicionado a la presencia o ausencia de varios factores de riesgo de fraude. Los factores de riesgo de fraude identificados en el modelo final incluían un sistema de control interno débil, un crecimiento rápido de la empresa, una rentabilidad relativa inadecuada o inconsistente, una gestión que sólo quiere lograr las proyecciones de ganancias de cualquier manera mientras miente a los auditores o es excesivamente evasiva, la condición de propiedad de la empresa (pública frente a privada), y el término de interacción entre un entorno de control débil y una actitud gerencial agresiva hacia la presentación de informes financieros.

### **2.1.5.2. Modelo de regresión logística**

El modelo logístico es un modelo lineal generalizado que se utiliza para la regresión binomial en el que las variables predictoras pueden ser numéricas o categóricas (Spathis, 2002; Yeh & hui Lien, 2009). Se utiliza principalmente para resolver los problemas causados por los seguros y el fraude corporativo. Algunas de las investigaciones han sugerido un modelo basado en la regresión logística para predecir la presencia de fraude en los estados financieros. El método estadístico de regresión logística puede detectar estados financieros falsificados de manera eficiente (Spathis, 2002)

El modelo Logit se presenta a continuación:

$$G(z) = \frac{e^z}{1 + e^z}$$

Donde  $z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$ , donde k es el número de variables a manejar

### **2.1.5.3. Árboles de decisión**

Los árboles de decisión son herramientas de apoyo a la decisión predictiva que crean mapas desde las observaciones hasta las posibles consecuencias. Estos árboles pueden plantarse mediante algoritmos basados en el aprendizaje por máquina como el ID3, el CART y el C4.5. Las predicciones se representan por hojas y las conjunciones de los rasgos por ramas. Los árboles de decisión se utilizan comúnmente en tarjetas de crédito, seguros de automóviles y fraudes corporativos (Han et al., 2011)

En un estudio el clasificador Naïve Bayes superó al clasificador convencional. La eficiencia de la predicción de los datos sobre el fraude financiero fue mayor, ya que no hubo falsos positivos y los falsos negativos fueron relativamente bajos (Phua et al., 2010)

Un árbol de decisión (DT) es un árbol estructurado, una herramienta de apoyo a la decisión, donde cada nodo representa una prueba y cada rama representa las posibles consecuencias. De esta manera, el modelo predictivo intenta dividir las observaciones en subgrupos mutuamente exclusivos y se utiliza para la extracción de datos y las tareas de machine learning. Los árboles de decisión son herramientas predictivas de apoyo a la decisión que crean mapas a partir de observaciones a las posibles consecuencias. Estos árboles pueden ser plantados a través de algoritmos basados en el machine learning. Las predicciones están representadas por las hojas las conjunciones de características por ramas. Los árboles de decisión son comúnmente usados en tarjetas de crédito, seguros de automóviles, y fraude corporativo. Para identificar y predecir el impacto de estados financieros fraudulentos, clasificación y regresión. (Sharma & Panigrahi, 2013)

### **2.1.5.4. Vector Support Machine**

Tiene como objetivo encontrar un hiperplano en un espacio N-dimensional (N - el número de características) que clasifica claramente los puntos de datos. Para separar las dos clases de puntos de datos, hay muchos hiperplanos posibles que podrían elegirse. Nuestro objetivo es encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre puntos de datos de ambas clases. Maximizar la distancia del margen proporciona cierto refuerzo para que los puntos de datos futuros puedan clasificarse con

más confianza. Este modelo es muy utilizado en aplicaciones de detección de fraude contable (Sharma & Panigrahi, 2013)

#### **2.1.5.5. Procesos gaussianos**

Los procesos gaussianos son una herramienta poderosa en la caja de herramientas de aprendizaje automático. Nos permiten hacer predicciones sobre nuestros datos incorporando conocimientos previos. Su área de aplicación más obvia es ajustar una función a los datos. Esto se denomina regresión y se utiliza, por ejemplo, en robótica o en predicciones de series de tiempo. Pero los procesos gaussianos no se limitan a la regresión, sino que también pueden extenderse a tareas de clasificación y agrupamiento. Para un conjunto dado de puntos de entrenamiento, existen potencialmente infinitas funciones que se ajustan a los datos. Los procesos gaussianos ofrecen una elegante solución a este problema al asignar una probabilidad a cada una de estas funciones. La media de esta distribución de probabilidad representa entonces la caracterización más probable de los datos. Además, el uso de un enfoque probabilístico nos permite incorporar la confianza de la predicción en el resultado de la regresión (Sadgali et al., 2019).

#### **2.1.5.6. Redes bayesianas**

Las redes bayesianas (BBN) representa un conjunto de variables aleatorias y sus independencias condicionales. La red de creencias bayesiana se utiliza en el desarrollo de modelos para la detección de fraudes en tarjetas de crédito, seguros de automóviles y empresas. La investigación realizada por (Kirkos et al., 2007) describió que el modelo de red de creencias bayesiana clasificó correctamente el 90,3% de la muestra de validación para la detección de fraudes. La red de creencias bayesiana superó los métodos de red neuronal y de árbol de decisión y logró una extraordinaria precisión de clasificación.

#### **2.1.5.7. Redes neuronales**

Las redes neuronales son herramientas de modelado de datos estadísticos no lineales que se inspiran en la funcionalidad del cerebro humano utilizando un conjunto de nodos interconectados (Ghosh & Reilly, 1994)

Las redes neuronales se aplican ampliamente en la clasificación y agrupación, y sus ventajas son las siguientes. En primer lugar, son adaptables; en segundo lugar, pueden

generar modelos robustos; y en tercer lugar, el proceso de clasificación puede modificarse si se establecen nuevos pesos de entrenamiento. Las redes neuronales se aplican principalmente a los fraudes con tarjetas de crédito, seguros de automóviles y fraudes empresariales (Sharma & Panigrahi, 2013)

En la literatura se describe que las redes neuronales pueden utilizarse como herramienta de detección de fraudes financieros. El modelo de clasificación de fraude de la red neural que emplea datos financieros endógenos creados a partir del patrón de comportamiento aprendido puede ser aplicado a una muestra de prueba. Las redes neuronales pueden ser usadas para predecir la ocurrencia de fraude corporativo a nivel gerencial (Cerullo & Cerullo, 1999)

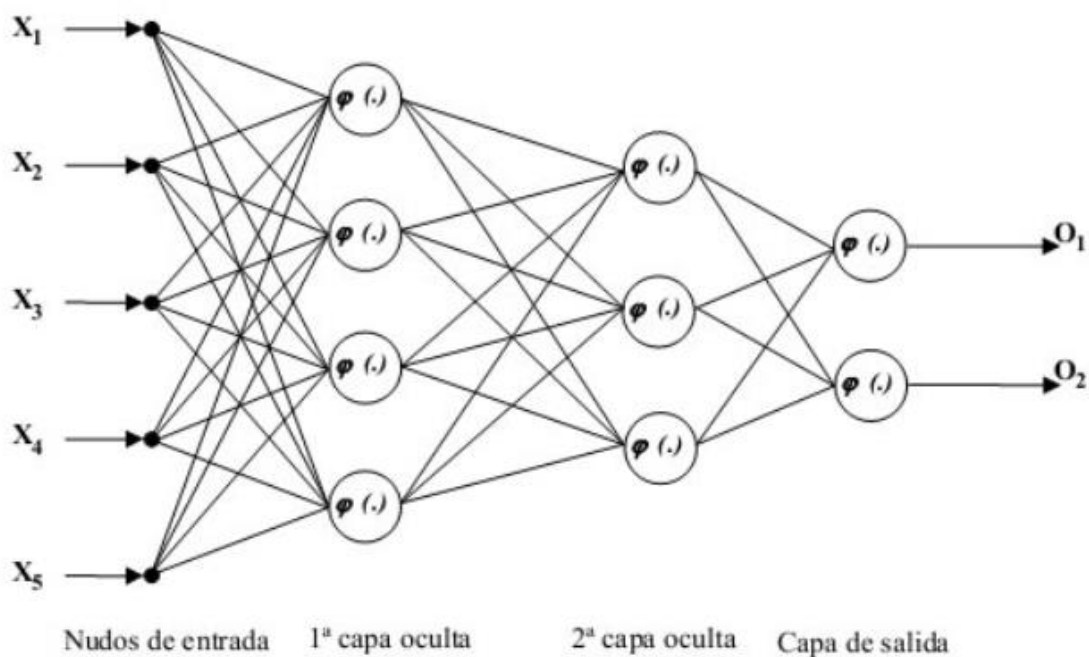
Es un modelo computacional que se inspira en la manera en que las redes neuronales del cerebro humano procesan la información (Tong, 2019)

La unidad básica en la neurona recibe datos de una fuente externa o de otros nodos para hallar una respuesta. Cada entrada tiene un peso asociado  $w$ , que se le asigna debido a su importancia con otras entradas.

El nodo aplica una función  $f$  la suma ponderada de sus entradas (Tong, 2019)

$$Y = f(\omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \dots + \omega_n x_n + \text{sesgo})$$

La función principal del sesgo es proporcionar a cada nodo un valor constante entrenable (además de las entradas normales que recibe el nodo). La función  $f$  no es lineal y se denomina Función de activación. El propósito de la función de activación es introducir la no linealidad en la salida de una neurona.



**Figura 2.** Arquitectura básica de una red neuronal

### 2.1.6. Métricas de desempeño

Una vez que haya construido su modelo, la pregunta más importante que surge es ¿qué tan bueno es su modelo? Por lo tanto, la evaluación de su modelo es la tarea más importante en el proyecto que delinea qué tan buenas son sus predicciones (Tong, 2019).

Una matriz de confusión es una tabla que se utiliza a menudo para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba cuyos valores reales se conocen (Tong, 2019)

**Verdadero Positivo:** Estos son los valores positivos correctamente predichos, lo que significa que el valor de la clase real es sí y el valor de la clase pronosticada es también sí. Por ejemplo, si el valor real de la clase indica que el pasajero sobrevivió y la clase prevista le dice lo mismo.

**Verdadero Negativo:** Estos son los valores negativos correctamente predichos, lo que significa que el valor de la clase real es no y el valor de la clase predichos también es no. Por ejemplo, si la clase real dice que este pasajero no sobrevivió y la clase prevista le dice lo mismo.

**Falso Positivo y Falso Negativo,** estos valores ocurren cuando su clase real contradice a la clase predicha.

**Falso Positivo:** Cuando la clase real es no y la clase pronosticada es sí. Por ejemplo, si la clase real dice que este pasajero no sobrevivió, pero la clase pronosticada le dice que este pasajero sobrevivirá.

**Falso Negativo:** Cuando la clase real es sí, pero la clase pronosticada en no. Por ejemplo, si el valor real de la clase indica que el pasajero sobrevivió y la clase pronosticada le indica que el pasajero morirá.

De acuerdo con (Tong, 2019), las métricas de desempeño más frecuentes son la exactitud, la precisión y la sensibilidad, ya que son relativamente sencillas de calcular e interpretar.

#### 2.1.6.1. Exactitud

La exactitud es la medida de rendimiento más intuitiva y es simplemente una relación entre la observación correctamente predicha y las observaciones totales. Uno puede pensar que, si tenemos una alta exactitud, entonces nuestro modelo es el mejor. Sí, la exactitud es una gran medida, pero sólo cuando se dispone de conjuntos de datos simétricos en los que los valores de falsos positivos y falsos negativos son casi los mismos. Por lo tanto, usted tiene que mirar otros parámetros para evaluar el rendimiento de su modelo.

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

#### 2.1.6.2. Precisión

La precisión es la relación entre las observaciones positivas correctamente predichas y el total de observaciones positivas predichas. La pregunta de que esta respuesta métrica es de todos los pasajeros que fueron etiquetados como sobrevivientes, ¿cuántos realmente sobrevivieron? La alta exactitud se relaciona con la baja tasa de falsos positivos.

$$\text{Precisión} = \frac{VP}{VP + FP}$$

### 2.1.6.3. Sensibilidad o Recall

Es el número de elementos identificados correctamente como positivos del total de positivos verdaderos.

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

### 2.1.6.4. Puntuación F1

La puntuación F1 es una medida de precisión en la evaluación de modelos de clasificación binaria o multi-etiqueta. Es especialmente útil en situaciones donde las clases no están balanceadas. La puntuación F1 se define como el promedio armónico de la precisión y la sensibilidad (o recall). Es decir, combina tanto la precisión (qué porcentaje de las predicciones positivas del modelo son realmente correctas) como la sensibilidad (qué porcentaje de los positivos reales fueron identificados correctamente por el modelo). La puntuación F1 se calcula de la siguiente manera:

$$F1 = \frac{1}{\frac{1}{\text{Precisión}} + \frac{1}{\text{Sensibilidad}}}$$

Un valor de F1 de 1 indica una precisión y sensibilidad perfectas, mientras que un valor de 0 indica que el modelo no tiene ni precisión ni sensibilidad. Por lo tanto, mientras más cercana esté la puntuación F1 a 1, mejor será el rendimiento del modelo en términos de precisión y sensibilidad.



## **2.2. Estado del arte**

A continuación, se revisan las investigaciones relacionadas a la aplicación de algoritmos de machine learning en la recaudación de impuesto predial.

### **2.2.1. Antecedentes internacionales**

La investigación de (Perbendaharaan et al., 2022) tuvo como objetivo proponer una solución para utilizar los datos almacenados por la Dirección General de Tributos como apoyo a la toma de decisiones y descubrir el potencial fiscal inexplorado y presentar modelos de análisis predictivo para predecir qué contribuyentes tienen más probabilidades de pagar lo que les corresponde. Para ello se utilizaron tres modelos de aprendizaje automático: regresión logística, bosque aleatorio y árbol de decisión. Se analizaron 5.562 muestras de datos de ingresos fiscales potenciales con ocho predictores: valor nominal de los datos de activación, distancia a la oficina de impuestos, tipo de contribuyente, soporte del informe fiscal, tipo de impuesto, estado del informe, año de registro del contribuyente y cobertura de la zona. Se evaluó el rendimiento de cada modelo y se determinó cuál proporcionó el mejor rendimiento de predicción. Se analizó la ponderación resultante de cada atributo para determinar la variable de mayor importancia en la predicción del potencial de ingresos fiscales. Como resultados el modelo de bosque aleatorio proporcionó el mejor rendimiento de predicción. El estado del informe fiscal fue la variable de mayor importancia en la predicción del potencial de ingresos fiscales. En conclusión, el enfoque de análisis predictivo puede proporcionar a los administradores tributarios una poderosa herramienta para aumentar la eficiencia del trabajo, combatir la evasión fiscal y ofrecer un mejor servicio al cliente.

La investigación de (Didimo et al., 2020), tuvo como objetivo presentar un enfoque novedoso llamado MALDIVE, que utiliza un modelo de red para apoyar a las administraciones tributarias en la evaluación del riesgo fiscal para descubrir la evasión y elusión fiscal. Metodología: MALDIVE se basa en un modelo de red que describe varios tipos de relaciones entre los contribuyentes. Este enfoque combina varios métodos de análisis visual y minería de datos para identificar a los contribuyentes de riesgo. MALDIVE consta de un canal de 4 pasos que incluye la creación de una red social a partir de los datos de los contribuyentes, la identificación de un conjunto inicial de contribuyentes de riesgo mediante la aplicación de algoritmos de aprendizaje automático,

la ampliación del conjunto de contribuyentes de riesgo mediante una estrategia de difusión de información y la validación y refinamiento del conjunto de contribuyentes riesgosos mediante la inspección visual de la red. Resultados: Los autores discuten la efectividad del enfoque MALDIVE a través de análisis cuantitativos y estudios de casos realizados en datos reales en colaboración con la Agencia Tributaria italiana. Los resultados muestran que MALDIVE es un enfoque efectivo para la evaluación del riesgo fiscal y la identificación de contribuyentes de riesgo. Conclusiones: El enfoque MALDIVE presentado en este documento proporciona una herramienta valiosa para las administraciones tributarias en la lucha contra la evasión y elusión fiscal. Los resultados muestran que la combinación de análisis visual, minería de datos y aprendizaje automático puede ayudar a identificar a los contribuyentes de riesgo y mejorar la eficacia de la administración tributaria. MALDIVE puede ser utilizado por las administraciones tributarias para detectar y prevenir la evasión y elusión fiscal, aumentando así la eficiencia de la administración tributaria y mejorando la equidad del sistema fiscal.

El artículo de (Galindo & Tamayo, 2000) se centra en la evaluación de riesgos de intermediarios financieros y la importancia de encontrar predictores precisos del riesgo individual en las carteras de crédito de las instituciones. El objetivo principal del estudio es hacer un análisis comparativo de diferentes métodos de modelado estadístico y de aprendizaje automático de clasificación en un conjunto de datos de préstamos hipotecarios con la motivación de comprender sus limitaciones y potencial. Para lograr este objetivo, el estudio utiliza una metodología específica de modelado basada en el estudio de las curvas de error y técnicas de modelado de vanguardia para construir más de 9,000 modelos. Los resultados muestran que los modelos de árbol de decisiones de CART proporcionan la mejor estimación de incumplimiento con una tasa de error promedio del 8,31% para una muestra de entrenamiento de 2000 registros. Además, si hubiera más datos disponibles, aproximadamente 22 000 registros, se podría lograr una tasa de error potencial del 7,32%. En comparación, Neural Networks proporcionó los segundos mejores resultados con un error promedio del 11.00%, mientras que el algoritmo K-Nearest Neighbor tuvo una tasa de error promedio del 14.95%. Estos resultados superaron al algoritmo Probit estándar, que alcanzó una tasa de error promedio del 15,13%. Como conclusión, los autores discuten las posibilidades de utilizar este tipo de modelo predictivo preciso como ingredientes de modelos de riesgo institucionales y

globales. En resumen, el artículo destaca la importancia de la evaluación precisa de riesgos en los intermediarios financieros y muestra que los modelos de árbol de decisiones de CART son los más eficaces para este propósito, lo que puede mejorar significativamente la eficiencia del uso de recursos en el sector financiero.

### **2.2.2. Antecedentes nacionales**

La tesis de (Aceituno Rojo, 2019) tuvo como objetivo determinar el mejor modelo de Machine Learning para mejorar el nivel de asertividad en el otorgamiento de microcréditos. Metodología: Se realizó una revisión de la literatura relacionada con el tema de la investigación y se analizó el proceso de otorgamiento de microcréditos de la entidad. Se identificaron 34 variables independientes para el entrenamiento de los modelos de Machine Learning. Se prepararon los datos para el entrenamiento de los modelos y se evaluaron a través de diversas métricas, como Accuracy, Precision, Recall, F1 Score y AUC ROC. Resultados: Como resultado se obtuvo la identificación de un total de 34 variables independientes para el entrenamiento de los modelos a fin de determinar si un crédito debe ser otorgado o no, así mismo se determinó que el modelo más asertivo es Artificial Neural Networks con un nivel de asertividad 93.72 % en comparación con los modelos Regresión Logística 86.07 %, Random Forest 66.35 %, Support Vector Machine 84.44 %, Decision Tree 88.80% y k-Nearest Neighbors 65.98 %, los cuales también ofrecen un nivel alto de asertividad. Conclusiones: La investigación concluye que el uso de Machine Learning puede mejorar el nivel de asertividad en el otorgamiento de microcréditos. El modelo más asertivo es Artificial Neural Networks, pero también se pueden considerar otros modelos con un alto nivel de asertividad. La identificación de las variables independientes es crucial para el entrenamiento de los modelos.

En el trabajo de Colonia, (2012). Titulado “Validez de pronóstico del modelo Credit-Scoring de una entidad microfinanciera” se presenta un modelo Credit-Scoring en una entidad microfinanciera, el cual permita calificar los solicitantes de un microcrédito. El modelo propuesto se basa en el modelo de regresión logística. La evaluación del modelo incluye el ajuste medido por la estadística Hosmer-lemeshow, poder predictivo medido por la  $R^2$ , poder discriminatorio medido por la curva COR y área bajo la curva COR. Se utilizó la información de clientes de la CMACT durante el periodo de agosto 2008 a julio del 2009 como base para desarrollar el modelo; y del periodo agosto a octubre

del 2009 para probar la validez del pronóstico. Para la clasificación del perfil del cliente se utilizaron 23 variables que incluyen variables sociodemográficas, variables de las características del crédito y variables de historial crediticio. Para clasificar el nivel de riesgo se utilizó la clasificación de la SBS (Normal, CPP, Deficiente, Dudoso y Pérdida). Como conclusión del estudio se encontró que el modelo es confiable, presentando un  $R^2$  de 83% y una tasa global de aciertos de 82%.

En el trabajo de Rodríguez, (2018). Titulado Modelo Credit-Scoring para obtener la probabilidad de impago de créditos personales futuros en la empresa servicios financieros grupo Buro, Huaraz 2018, se presenta un modelo Credit-Scoring utilizado como una herramienta complementaria para la decisión final en el otorgamiento o rechazo de la solicitud de crédito. Se trata de una metodología que pronostica el riesgo de incumplimiento de pagos futuros, a través de la designación de un score o puntaje, basado en datos crediticios, financiero, laboral y sociodemográfico referente a los clientes. Los datos para tal fin son de carácter socio demográfico, laboral, financiero y crediticio, datos internos es decir de la misma empresa, basados en su experiencia. De esta manera se pudo estimar la probabilidad de casos buenos y malos que representan los atributos o características de cada variable y a partir de estas probabilidades se genera un puntaje, es decir un score para indicar un nivel de riesgo. Para caracterizar el perfil del cliente se escogieron 16 variables y para estimar el nivel de riesgo se usó una variable dicotómica (pago o impago). El modelo Credit-Scoring elaborado tiene el 76% de predicción, todas las variables incluidas presentan signo coherente de influencia en la variable explicativa, las pruebas de Significación de chi-cuadrado del modelo cuyos valores son menores a 0,05 lo cual indica que el modelo ayuda a explicar el evento, además de R-cuadrado de Cox y Snell, y R-cuadrado de Nagelkerke indican que las variables independientes explican significativamente la variable dependiente. La probabilidad de impago obtenida a través de la regresión logística se transformó en una escala de 0 a 1000 puntos o score, para crear una herramienta cuantitativa que nos indique si se debe conceder o no el crédito, para ello la entidad financiera determino que el punto de corte es 430, es decir si el score es  $> 430$  se otorga el crédito, en caso contrario si  $\text{score} \leq 430$  se rechaza la solicitud.

### III. METODOLOGÍA

#### 3.1. Tipo y nivel de investigación

La presente investigación se clasifica como una investigación aplicada, dado que se orienta hacia la resolución de un problema práctico específico: predicción de pagadores del impuesto predial en la ciudad de Tingo María. La investigación aplicada se enfoca en la utilización de teorías, principios y técnicas con el objetivo de abordar y resolver problemas concretos del mundo real. En este contexto, se implementan algoritmos de Machine Learning para identificar patrones de fraude.

Respecto al nivel de investigación, este estudio se categoriza como una investigación explicativa. Este tipo de investigación busca comprender las causas y los efectos de un fenómeno. En este caso, se procura entender cómo diferentes algoritmos de Machine Learning pueden influir en la capacidad para detectar el fraude en el impuesto predial. La investigación explicativa representa el nivel más profundo de investigación, ya que no solo describe o explora un fenómeno, sino que también busca comprender las razones de su ocurrencia.

#### 3.2. Población y muestra

En la ciudad de Tingo María existen 12821 contribuyentes y de acuerdo con la base de datos de La Municipalidad Provincial de Leoncio Prado, 4365 pagan el impuesto predial de manera puntual, 187 pagan con retrasos y 8269 no pagan.

El tamaño de muestra se determinará usando la fórmula de muestreo simple aleatorio:

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1 - p)}{(N - 1) \cdot e^2 + Z^2 \cdot p \cdot (1 - p)}$$

Donde  $Z^2$  es la puntuación  $z$  para un nivel de confianza  $1 - \alpha$ ,  $p$  es la probabilidad de éxito,  $e$  es el margen de error en las mediciones y  $N$  es el tamaño total de la población. Para  $Z^2=1.96$ ,  $\alpha = 0.05$ ,  $e = 0.05$ ,  $p = 0.5$  y  $N = 12821$  el tamaño de muestra es  $372.07 \cong 372$  encuestas.

Considerando que para lograr una mayor equidad en el aprendizaje de las categorías es necesario que la muestra sea equidistribuida, se considerará un aproximado

de 124 personas por cada categoría de contribuyente (PAGAN, IMPUNTUALES, NO PAGAN)

### **3.3. Hipótesis**

#### **3.3.1. General**

Existe diferencia en el desempeño de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

#### **3.3.2. Específicas**

Existe diferencia en la exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

Existe diferencia en la precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

Existe diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019

### **3.4. Variables de investigación**

#### **3.4.1. Variable dependiente**

Desempeño de modelos de Machine Learning

#### **3.4.2. Variable independiente**

Modelos de Machine Learning

### 3.5. Operacionalización de variables

Variable	Dimensiones	Definición
Desempeño de modelos de Machine Learning	F1	La puntuación F1 se define como el promedio armónico de la precisión y la sensibilidad (o recall). Es decir, combina tanto la precisión (qué porcentaje de las predicciones positivas del modelo son realmente correctas) como la sensibilidad (qué porcentaje de los positivos reales fueron identificados correctamente por el modelo).
	Exactitud	La exactitud es la proporción de predicciones correctas hechas por el modelo sobre todas las predicciones realizadas.
	Precisión	La precisión es la proporción de identificaciones positivas que fueron realmente correctas.
	Sensibilidad (Recall)	Sensibilidad o Recall se refiere a la proporción de positivos reales que fueron identificados correctamente.
Modelos de Machine Learning	Decisión Tree	Los Árboles de Decisión son un tipo de modelo que utiliza una estructura de árbol para tomar decisiones.
	Support Vector Machine	Las Máquinas de Vectores de Soporte son un conjunto de algoritmos de aprendizaje supervisado que se utilizan para problemas de clasificación y regresión.
	K - Nearest Neighbors	K - Nearest Neighbors es un algoritmo que almacena todos los casos disponibles y clasifica los nuevos casos en función de una medida de similitud.
	Naive Bayes	Naive Bayes es un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes con la suposición "ingenua" de independencia condicional entre cada par de características.
	Ensemble	Los métodos de Ensemble son técnicas que crean múltiples modelos y luego combinan para producir una salida mejorada.
	Redes Neuronales	Las Redes Neuronales son un conjunto de algoritmos modelados a partir del cerebro humano, diseñados para reconocer patrones.

### **3.6. Herramientas de recolección de datos**

Para la recolección de datos de esta investigación, se empleó una herramienta de encuesta telefónica. Esta herramienta fue seleccionada debido a su eficacia para obtener información directa y actualizada de los contribuyentes. La base de datos proporcionada por la Municipalidad Provincial de Leoncio Prado sirvió como punto de partida para identificar a los contribuyentes y clasificarlos en tres categorías: pagadores puntuales, pagadores tardíos y no pagadores.

El tamaño de la muestra se calculó para garantizar una representación adecuada de la población de contribuyentes. Sin embargo, debido a la falta de disponibilidad de algunos contribuyentes, fue necesario contactar a un número mayor de individuos que el inicialmente previsto.

La encuesta telefónica, diseñada específicamente para este estudio, permitió recopilar datos relevantes sobre el comportamiento de pago de los impuestos prediales. Esta herramienta de recolección de datos demostró ser eficaz para obtener la información necesaria para el análisis posterior.

Es importante destacar que todas las herramientas y métodos de recolección de datos utilizados en esta investigación se seleccionaron con el objetivo de garantizar la precisión y la relevancia de los datos recopilados.

### **3.7. Metodología**

#### **3.7.1. Análisis descriptivo**

La información que se utilizará para entrenar los modelos de Machine Learning será información de tipo socioeconómica y de percepción del contribuyente, de manera que a partir de este perfil se pueda predecir la categoría de contribuyente (PAGAN, IMPUNTUALES, NO PAGAN), para ello, a partir de la base de datos de los contribuyentes y su categoría de pago de impuesto predial (PAGAN, IMPUNTUALES, NO PAGAN), se le realizará una llamada telefónica y se realizará una encuesta asistida en donde se aplicará un cuestionario de preguntas cerradas como se muestra en el ANEXO 1.

El análisis descriptivo consistirá en un conteo de las respuestas realizadas por los contribuyentes por cada categoría, esto se representará en un mapa de calor.



Posteriormente se realizará la prueba de  $\chi^2$  para determinar si esos indicadores de información socioeconómica o de percepción influyen en la categoría de contribuyente (PAGAN, IMPUNTUALES, NO PAGAN). Ambos análisis estadísticos (mapa de calor y prueba de  $\chi^2$ ) se llevarán a cabo en el software de prueba Matlab R2023a, donde se utilizó una licencia Trial- Individual gratuita de prueba (Licencia 11355186).

### **3.7.2. Métricas de desempeño**

#### **3.7.2.1. Exactitud**

Para determinar la exactitud, en primer lugar, se construirán las matrices de confusión de cada modelo, los cuales son:

- Decisión Tree
- Support Vector Machine
- K - Nearest Neighbors
- Naive Bayes
- Ensemble
- Redes Neuronales

Una vez obtenidas las matrices de confusión de cada modelo, se aplicará la siguiente fórmula para determinar la exactitud.

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

Los modelos serán entrenados a partir de los datos de la encuesta mencionada anteriormente, el método de evaluación será la validación cruzada k-fold con un valor de  $k = 5$ , el entrenamiento se llevará a cabo en el software de prueba Matlab R2023a, donde se utilizó una licencia Trial- Individual gratuita de prueba (Licencia 11355186) en la App Classification Learner en donde se optimizarán los hiperparámetros del modelo para 30 iteraciones. Las variables predictoras seleccionadas serán aquellas que resultaron con influencia significativa en la prueba de  $\chi^2$  para un valor de  $\alpha = 0.01$ .

#### **3.7.2.2. Precisión**

La precisión se determinará mediante la siguiente fórmula:

$$\text{Presición} = \frac{VP}{VP + FP}$$

Donde VP representa a la cantidad de categorías predichas correctamente para cada categoría de pagador (PAGAN, IMPUNTUALES, NO PAGAN) y VF representa la cantidad de predicciones erróneas de las categorías de pagador del contribuyente.

Posteriormente se compararán las precisiones de los modelos de machine learning mencionados anteriormente y se propondrá una propuesta de caso para la aplicación del algoritmo más preciso. La propuesta de caso consistirá en la identificación específica de una muestra para una determinada restricción de condiciones que responda a la pregunta ¿qué porcentaje de las categorías predichas son las correctas?, en este caso se necesitará al modelo más preciso. Finalmente se agruparán en una tabla las precisiones de todos los modelos para cada categoría de pagador (PAGAN, IMPUNTUALES, NO PAGAN).

### 3.7.2.3. Sensibilidad o Recall

La sensibilidad se determinará mediante la siguiente fórmula:

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

Donde VP representa a la cantidad de categorías predichas correctamente para cada categoría de pagador (PAGAN, IMPUNTUALES, NO PAGAN) y FN representa la cantidad de predicciones erróneas de las que no eran las categorías de pagador del contribuyente.

Posteriormente se compararán las precisiones de los modelos de machine learning mencionados anteriormente y se propondrá una propuesta de caso para la aplicación del algoritmo más sensible. La propuesta de caso consistirá en la identificación total de los tipos de pagadores, la predicción responderá a la ¿qué porcentaje de las categorías somos capaces de identificar del total existente en una muestra?, en este caso se necesitará al modelo más sensible. Finalmente se agruparán en una tabla las precisiones de todos los modelos para cada categoría de pagador (PAGAN, IMPUNTUALES, NO PAGAN).

### 3.7.2.4. Puntuación F1

La puntuación F1 puede ser una métrica útil para evaluar el rendimiento de un modelo de aprendizaje automático, especialmente en problemas de clasificación donde las clases están desequilibradas. En este sentido se tomará la Puntuación F1 para evaluar el desempeño general de los modelos de machine learning:

$$F1 = \frac{2}{\frac{1}{\text{Presición}} + \frac{1}{\text{Sensibilidad}}}$$

Para interpretar la puntuación F1 en este caso, se necesita entender qué representan las tres categorías. **PAGADORES PUNTUALES:** El modelo tiene una buena puntuación F1 para esta categoría si identifica correctamente a la mayoría de los pagadores puntuales, y no clasifica erróneamente a muchos no pagadores puntuales como pagadores puntuales. **PAGADORES IMPUNTUALES:** El modelo tiene una buena puntuación F1 para esta categoría si identifica correctamente a la mayoría de los pagadores impuntuales, y no clasifica erróneamente a muchos no pagadores impuntuales como pagadores impuntuales. **PAGADORES TARDÍOS:** El modelo tiene una buena puntuación F1 para esta categoría si identifica correctamente a la mayoría de los pagadores tardíos, y no clasifica erróneamente a muchos no pagadores tardíos como pagadores tardíos.

La puntuación F1 promedio proporcionará una idea general de cuán bien está funcionando el modelo en todas las clases. Un valor de F1 cercano a 1 significa que el modelo tiene una buena precisión y sensibilidad en todas las clases. Un valor de F1 más cercano a 0 sugiere que el modelo tiene problemas para identificar correctamente las clases o que está haciendo muchas predicciones incorrectas.

### 3.7.3. Diferencias en el desempeño

Para identificar diferencias significativas entre los modelos de aprendizaje automático evaluados, se llevó a cabo un Análisis de Varianza de Friedman. Esta prueba no paramétrica se seleccionó debido a su idoneidad para comparar más de dos grupos relacionados sin asumir normalidad en la distribución de los datos. Se estableció un nivel de significancia de  $\alpha = 0.05$  para todas las pruebas estadísticas.

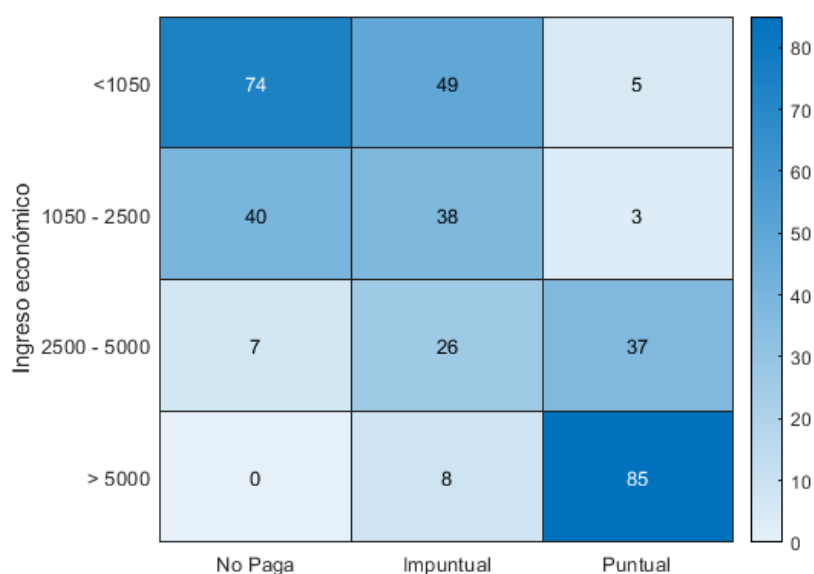
Una vez obtenidos los resultados del ANOVA de Friedman, se procedió con la prueba post hoc de Nemenyi para realizar comparaciones múltiples entre los modelos. La prueba de Nemenyi complementa el ANOVA de Friedman, permitiendo discernir qué modelos presentan diferencias significativas en sus rendimientos. Al igual que con el ANOVA de Friedman, se utilizó un nivel de significancia de  $\alpha = 0.05$  para determinar la presencia de diferencias estadísticas significativas entre los pares de modelos.

## IV. RESULTADOS Y DISCUSIÓN

### 4.1. Descripción e influencia de características

#### Ingreso económico

Los resultados que se muestran en la **Figura 3** indican que para los contribuyentes clasificados como "No pagadores", la mayoría tiene un ingreso mensual familiar menor a S/1050. Mientras que para los contribuyentes clasificados como "Puntuales", la mayoría tiene un ingreso mensual familiar mayor a S/5000. La prueba de Chi-Cuadrado realizada sugiere que existe una relación estadísticamente significativa entre el ingreso económico y la clasificación de tipo de pagador del contribuyente. En otras palabras, el nivel de ingreso económico de los contribuyentes está relacionado con su capacidad para realizar pagos puntuales. Este análisis puede ser útil para comprender la situación económica de los contribuyentes y su impacto en su capacidad para cumplir con sus obligaciones financieras.

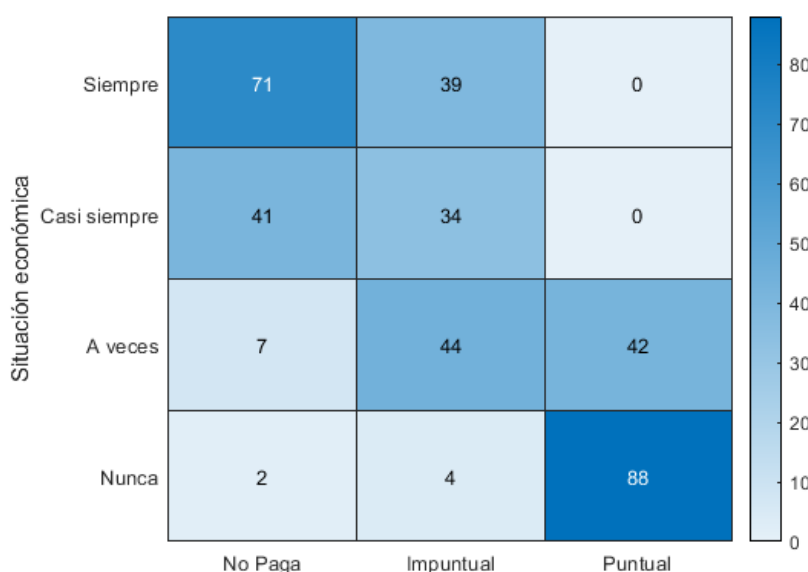


**Figura 3.** Ingreso económico de los diferentes tipos de contribuyentes

#### Situación económica

Los resultados de la **Figura 4** indican que para los contribuyentes clasificados como "No pagadores", la mayoría considera que su situación económica siempre impide

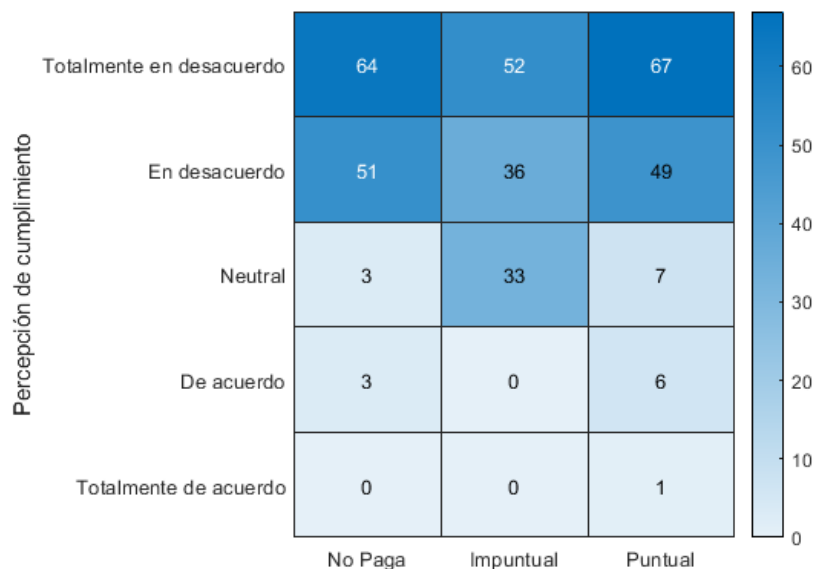
el pago del impuesto predial. Mientras que para los contribuyentes clasificados como "Puntuales", la mayoría considera que su situación económica nunca impide el pago del impuesto predial. La prueba de Chi-Cuadrado realizada demuestra que existe una relación estadísticamente significativa entre la situación económica y la clasificación de tipo de pagador del contribuyente. En otras palabras, la capacidad de los contribuyentes para realizar el pago del impuesto predial está relacionada con su situación económica.



**Figura 4.** Situación económica como causa de impago/impuntualidad/pago de tributos

### Percepción de cumplimiento

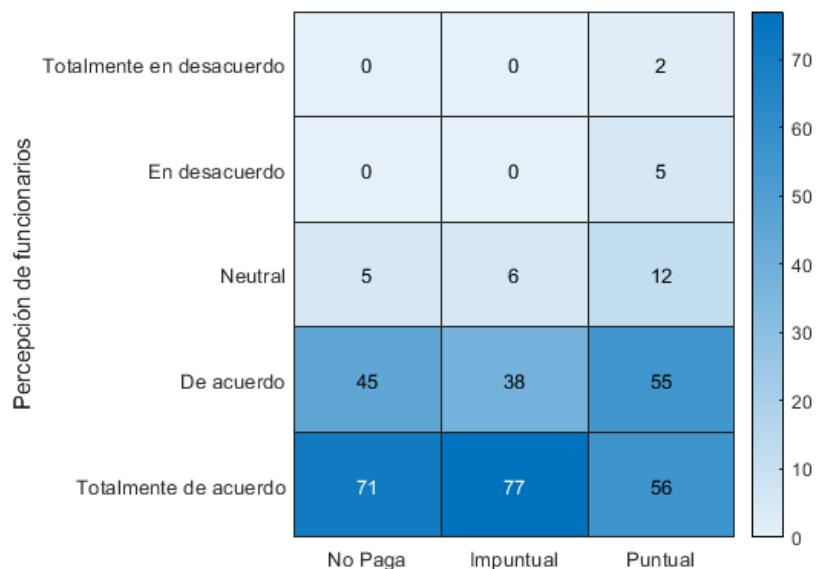
La **Figura 5** muestra los resultados de una encuesta sobre la percepción de cumplimiento de las autoridades de la municipalidad para solucionar los problemas de la provincia, las respuestas se dividen en cinco categorías: Totalmente en desacuerdo, En desacuerdo, Neutral, De acuerdo y Totalmente de acuerdo. Los resultados muestran que la mayoría de los contribuyentes clasificados como 'No pagadores', 'Impuntuales' y 'Puntuales' están en desacuerdo o totalmente en desacuerdo con la percepción de cumplimiento de las autoridades de la municipalidad. Además, se realizó una prueba de Chi-Cuadrado para determinar si la percepción de cumplimiento y la clasificación de tipo de pagador del contribuyente son dependientes. El resultado de la prueba indica que existe evidencia estadística altamente significativa para afirmar que la percepción de cumplimiento y la clasificación de tipo de pagador del contribuyente son dependientes.



**Figura 5.** Percepción de los contribuyentes del cumplimiento de las autoridades para solucionar problemas de la ciudad.

### Percepción de funcionarios

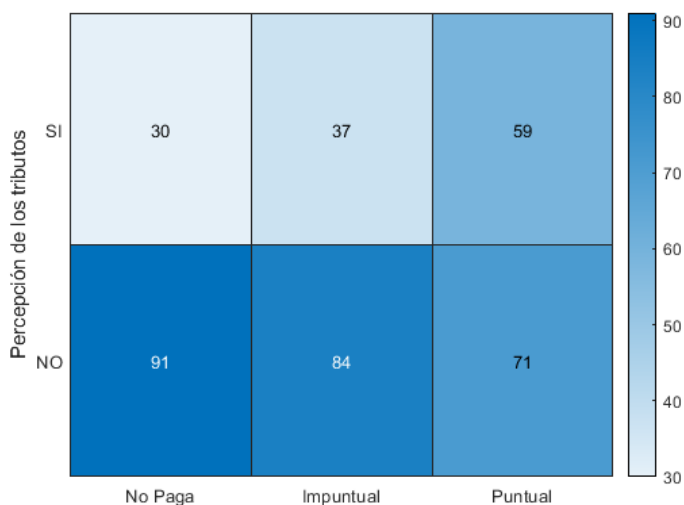
La **Figura 6** muestra los resultados de la encuesta sobre la percepción de corrupción de los funcionarios de la Municipalidad. Los resultados se presentan en forma de porcentajes y se dividen en cinco categorías: Totalmente en desacuerdo, En desacuerdo, Neutral, De acuerdo y Totalmente de acuerdo. Los resultados muestran que la mayoría de los contribuyentes clasificados como 'No pagadores', 'Impuntuales' y 'Puntuales' están de acuerdo o totalmente de acuerdo con la percepción de corrupción de los funcionarios de la Municipalidad. Además, se realizó una prueba de Chi-Cuadrado para determinar si la percepción de funcionarios y la clasificación de tipo de pagador del contribuyente son dependientes. El resultado de la prueba indica que existe evidencia estadística altamente significativa para afirmar que la percepción de funcionarios y la clasificación de tipo de pagador del contribuyente son dependientes.



**Figura 6.** Percepción de los contribuyentes con respecto a la corrupción de los funcionarios.

### Percepción de los tributos

En la **Figura 7** se muestran los resultados de la encuesta: el 75.21% de los contribuyentes clasificados como ‘No pagadores’ respondió que NO acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente. Para los contribuyentes clasificados como ‘Impuntuales’, el 69.42% respondió que NO acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente. Para los contribuyentes clasificados como ‘Puntuales’, el 54.62% respondió que NO acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente. Además, se realizó una prueba de Chi-Cuadrado para determinar si la percepción de los tributos y la clasificación de tipo de pagador del contribuyente son dependientes. El resultado de la prueba indica que existe evidencia estadística altamente significativa para afirmar que la percepción de los tributos y la clasificación de tipo de pagador del contribuyente son dependientes.

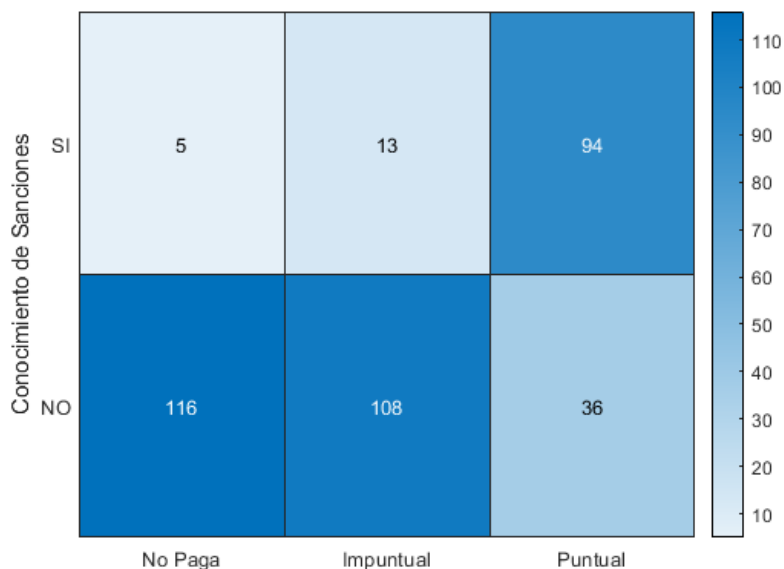


**Figura 7.** Percepción de los contribuyentes con respecto a la utilidad de los tributos.

### **Conocimiento de Sanciones**

En la **Figura 8** se puede observar que la mayoría de los contribuyentes no están conscientes de las sanciones o multas que pueden generar el no declarar información de sus predios. En particular, para los No pagadores, el 95.87% no son conscientes de las sanciones tributarias, mientras que, para los Impuntuales, el 89.26% no son consciente de las sanciones tributarias. Por otro lado, los Puntuales tienen un mayor conocimiento de las sanciones, con un 72.31% indicando que sí está consciente. Es importante destacar que al aplicar la prueba de Chi-Cuadrado, se encontró evidencia estadística altamente significativa que demuestra que el Conocimiento de Sanciones y la clasificación de tipo de pagador del contribuyente son dependientes. Es decir, hay una relación entre el conocimiento de sanciones y el tipo de pagador del contribuyente.

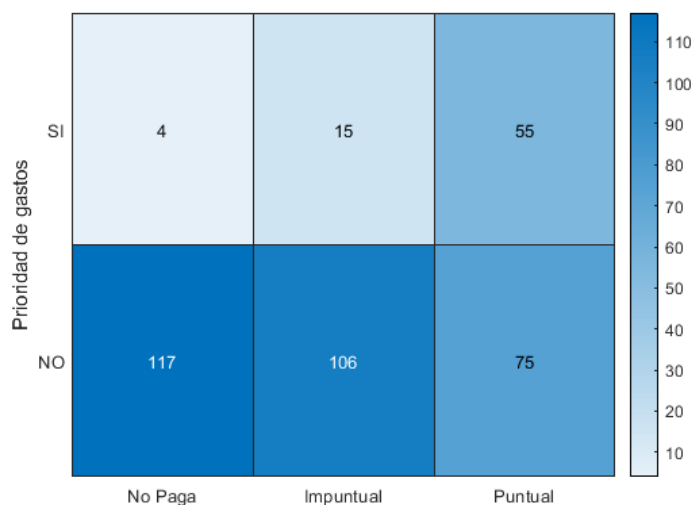




**Figura 8.** Conocimiento de los contribuyentes con respecto a las sanciones tributarias.

### **Prioridad de gastos**

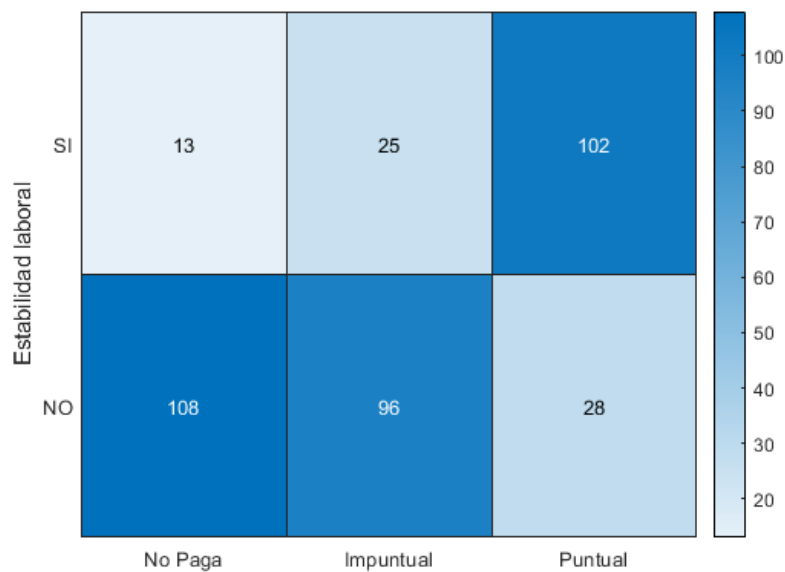
Los resultados de la **Figura 9** muestran que la mayoría de los contribuyentes clasificados como 'No pagadores' y 'Impuntuales' no consideran el pago del impuesto predial como primordial, con un porcentaje de respuesta negativa del 96.69 % y 87.6 %, respectivamente. Por otro lado, el 42.31 % de los contribuyentes clasificados como 'Puntuales' sí consideran el pago del impuesto predial como primordial. Además, al aplicar la prueba de Chi-Cuadrado se encontró una relación estadísticamente significativa entre la prioridad de gastos y la clasificación de tipo de pagador del contribuyente. Esto puede ser útil para la municipalidad al momento de diseñar estrategias para incentivar el pago del impuesto predial, ya que puede enfocarse en aquellos contribuyentes que consideran este pago como una prioridad en sus gastos.



**Figura 9.** Priorización de los contribuyentes con respecto a los gastos necesarios del hogar.

### **Estabilidad laboral**

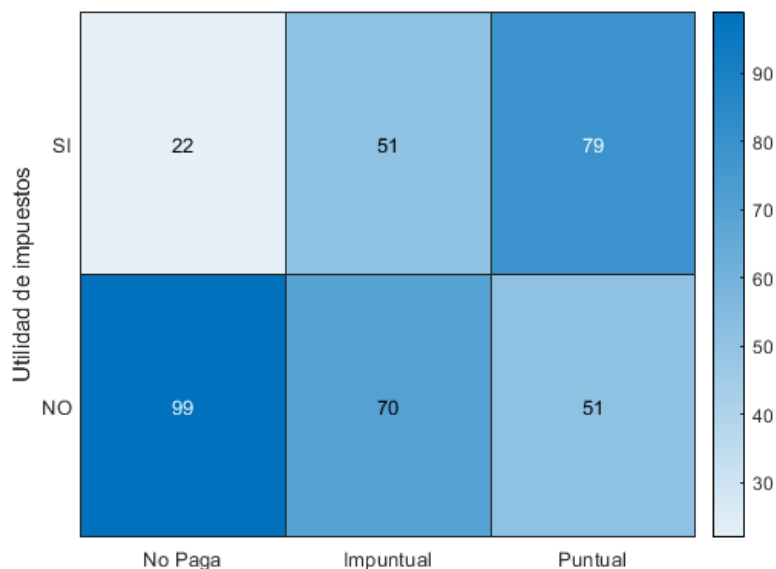
Los resultados de la **Figura 10** muestran una correlación entre la estabilidad laboral y la clasificación de tipo de pagador del contribuyente. En general, los contribuyentes puntuales tienen una mayor estabilidad laboral, con un 78.46% respondiendo afirmativamente a la pregunta, mientras que solo el 10.74% de los no pagadores lo hizo. Sin embargo, también hay un porcentaje considerable de contribuyentes impuntuales que tienen una estabilidad laboral mayor a 3 años, con un 20.66%. Esto podría indicar que la estabilidad laboral es un factor importante pero no determinante en la puntualidad del pago del impuesto predial.



**Figura 10.** Estabilidad laboral de los contribuyentes.

### Utilidad de impuestos

En el indicador 'Utilidad de impuestos' que se muestra en la **Figura 11** se observa que, en general, los contribuyentes que se clasifican como puntuales tienen una percepción más positiva sobre la calidad de los servicios públicos en la zona donde viven, con un 60.77% respondiendo que sí se brindan estos servicios. Por otro lado, los impuntuales tienen un 42.15% de respuestas afirmativas y los no pagadores solo un 18.18%. En cuanto a la prueba de Chi-cuadrado, con un valor calculado de 47.16 y un valor crítico de 0.02 se evidencia que la percepción de utilidad de impuestos y la clasificación de tipo de pagador del contribuyente están altamente relacionados estadísticamente.



**Figura 11.** Percepción de los contribuyentes con respecto a la utilidad de los tributos.

De acuerdo con la Tabla 3, todas las características planteadas en la encuesta resultan altamente significativas en el tipo de pagador del contribuyente.

**Tabla 2.** Variables socioeconómicas y de percepción que influyen en el tipo de pagador del contribuyente

Característica	p-valor*
Ingreso económico	< 0.01
Situación económica	< 0.01
Percepción de cumplimiento	< 0.01
Percepción de funcionarios	< 0.01
Percepción de los tributos	< 0.01
Conocimiento de Sanciones	< 0.01
Prioridad de gastos	< 0.01
Estabilidad laboral	< 0.01
Utilidad de impuestos	< 0.01

\* Valores de significancia de la prueba de Chi - Cuadrado

#### 4.2. Evaluación de la Exactitud

Los modelos de predicción mostraron diferentes niveles de exactitud. El modelo de Árbol de Decisión tuvo una exactitud del 91.94%, mientras que el modelo de Máquina de Vectores de Soporte alcanzó el 90.59%. El modelo de Vecinos más Cercanos (KNN) tuvo una exactitud del 97.04%, superando al modelo de Bayes Ingenuo que logró el 90.86%. El modelo de Conjunto tuvo una exactitud del 92.2%, y finalmente, el modelo

de Redes Neuronales logró una exactitud del 97.01%. De todos estos, los modelos KNN y de Redes Neuronales fueron los que tuvieron la mayor exactitud, con 97.04% y 97.01% respectivamente.

En comparación con la tesis de (Aceituno Rojo, 2019), se observan algunas diferencias en los resultados obtenidos en esta investigación. El modelo SVM en la tesis de Aceituno obtuvo una exactitud del 88.81%, mientras que, en esta investigación, se obtuvo una exactitud del 90.59%. Por otro lado, el modelo Redes Neuronales en la tesis de Aceituno tuvo la mayor exactitud con un 93.72%, mientras que, en esta investigación, este modelo tuvo una exactitud aun mayor (97.01%). En cuanto al modelo Decision Tree, en la tesis de Aceituno este modelo tuvo una exactitud del 92.3%, mientras que, en esta investigación, este modelo tuvo una exactitud ligeramente menor (91.94%). Finalmente, en la tesis de Aceituno, el modelo KNN tuvo una exactitud del 65.98%, mientras que, en esta investigación, este modelo obtuvo una exactitud muy alta del 97.04%. Cabe destacar que mientras en la tesis de Aceituno se utilizaron 34 variables independientes, en esta investigación solo se emplearon 11 variables socioeconómicas. Es posible que el exceso de variables en la tesis de Aceituno haya ocasionado que algunas variables no estuvieran muy correlacionadas con el resultado, mientras que, en esta investigación, la selección de un conjunto más reducido de variables haya permitido obtener mejores predicciones a pesar de tener menos variables de entrada.

En la investigación de (López et al., 2019), se evaluó la evasión del Impuesto sobre la Renta de las Personas Físicas (IRPF) en España, mediante el uso de herramientas predictivas avanzadas de Machine, mediante la aplicación de modelos de redes neuronales de Perceptrón Multicapa (MLP), la investigación encontró que la precisión del modelo de 84.19%, aunque un poco inferior a la exactitud de los modelos encontrados en este estudio (Exactitud  $\geq 90\%$ ), representa un buen rendimiento. La exactitud del modelo de redes neuronales de esta tesis es de 97.01%, lo que representa superioridad al estudio anterior citado, esta diferencia en la exactitud podría deberse al tipo de información de entrada, en la investigación de (López et al., 2019), se utiliza información de los estados financieros del contribuyente, sin embargo en esta investigación se recoge información socioeconómica y de la percepción del contribuyente, por lo que la información que utiliza esta tesis como predictores, puede ser que estén más direccionadas al comportamiento del cliente y reflejo más directamente su predisposición

al pago del impuesto predial. En el estudio de (Mesa et al., 2009), la exactitud de su modelo Logístico fue de 91.95%, sin embargo, cabe mencionar que en la muestra trabajada 352 personas SI cometieron fraude y 3727 NO cometieron fraude lo que representa un desbalance de las categorías, haciendo de esta exactitud poco confiable, en esta tesis la muestra trabajó proporcionalmente con cada categorías (NO PAGA = 127, IMPUNTUAL = 127, PUNTUAL = 118), mencionar también que el modelo posee exactitudes mayores al 90% y el modelo más exacto es el KNN con una exactitud de 97.04%.

En la investigación de (Baghdasaryan et al., 2022) se predijo el fraude fiscal a partir de variables como fraude histórico y auditoría, participación en los costos administrativos y actividad económica externa, aplicando el algoritmo de Machine Learning de Decisión Tree, se encontró una exactitud de 85.3%, exactitud cercana a la de esta tesis, 91.94% para el algoritmo de Decisión Tree, sin embargo, cabe mencionar que el desbalance de muestras es similar al de (Mesa et al., 2009). Los modelos de aprendizaje automático han demostrado ser instrumentos eficaces en la calificación de riesgo crediticio, y el presente estudio revela datos relevantes para contribuir a esta línea de investigación. En cuanto a la exactitud de los modelos, los hallazgos actuales muestran que los modelos de Vecinos más Cercanos (KNN) y Redes Neuronales exhiben un rendimiento superior con exactitudes de 97.04% y 97.01% respectivamente. Esto es notable, dada la prevalencia de la utilización de estos modelos en la calificación de riesgo crediticio. Los hallazgos sobre el modelo de Redes Neuronales corroboran la investigación de (Huang et al., 2004), quienes descubrieron que las Redes Neuronales superaban a las Máquinas de Vectores de Soporte en la calificación de riesgo crediticio. Este estudio revela resultados similares, con las Redes Neuronales exhibiendo una mayor exactitud que las Máquinas de Vectores de Soporte (90.59%). Los modelos de Árbol de Decisión, si bien presentan una alta exactitud de 91.94%, aún son superados por otros modelos, como el de Conjunto y Redes Neuronales. Este hecho está en línea con el estudio de (Lessmann et al., 2015), donde se resalta el rendimiento competitivo de los modelos de Árbol de Decisión, pero se reconoce su inferioridad frente a los métodos de ensamble. Uno de los hallazgos más interesantes de este estudio es el rendimiento sobresaliente del modelo KNN. Esto contrasta con el estudio de (Çiğşar & Ünal, 2019), donde el rendimiento de KNN fue inferior al de los modelos de Redes Neuronales y SVM. Este

hallazgo indica que podría haber factores específicos del conjunto de datos actual o del preprocesamiento de datos que favorecen el uso de KNN. En cuanto al modelo de Conjunto, los resultados corroboran el estudio de (Louzada et al., 2016), donde se muestra que los modelos de Conjunto tienden a superar a los modelos Bayes Ingenuos en términos de exactitud.

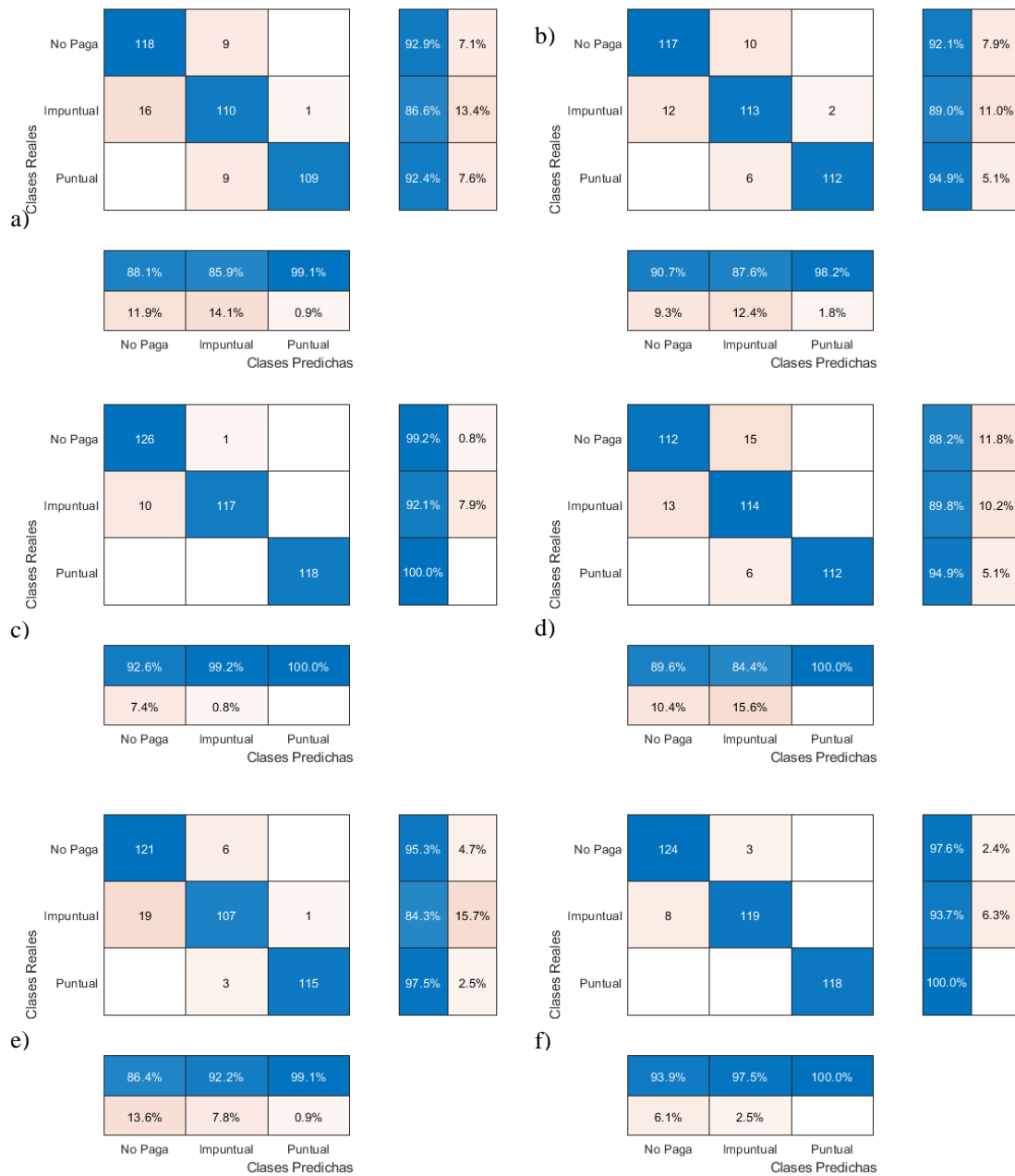
La Tabla 3 muestra los resultados del ANOVA de Friedman aplicado a la exactitud de diferentes algoritmos de aprendizaje automático, utilizando un nivel de significancia de  $\alpha = 0.05$ . El valor de Chi-cuadrado ( $\text{Chi}^2$ ) es de 42.86, con un p-valor menor que 0.01, lo que indica que existen diferencias estadísticamente significativas en la exactitud entre al menos algunos de los algoritmos evaluados. En la prueba post hoc de Nemenyi, que también se aplica a un nivel de significancia de  $\alpha = 0.05$ , se observan diferencias en los grupos estadísticos. El algoritmo con el mejor rendimiento en términos de exactitud es el K - Nearest Neighbors (KNN), con un promedio de exactitud del  $97.04 \pm 0.30$ , también incluye al algoritmo de Redes Neuronales (RNN) que presenta una exactitud de  $96.89 \pm 0.14$ , ambos clasificados en el grupo estadístico C, lo que demuestra que tienen una exactitud significativamente más alta en comparación con los otros modelos. Por otro lado, los modelos Support Vector Machine (SVM) y Naive Bayes, ambos en el grupo estadístico A, mostraron el menor rendimiento en términos de exactitud, con promedios de  $90.56 \pm 0.30$  y  $91.00 \pm 0.09$ , respectivamente. Esto sugiere que estos modelos comparten un nivel de exactitud similar y no son estadísticamente diferentes entre sí, pero sí son diferentes del modelo KNN, el cual se encuentra en un grupo estadístico distinto.

**Tabla 3.** ANOVA de Friedman de la Exactitud de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de  $\alpha = 0.05$ .

Fuente	SC	Gl	SM	$\text{Chi}^2$	p-valor
Algoritmos	150	5	30	42.86	< 0.01
Error	25	45	0.56		
Total	175	59			

Algoritmos	Promedio $\pm$ E.E.	Grupos estadísticos
Support Vector Machine (SVM)	$90.56 \pm 0.30$	A
Naive Bayes	$91.00 \pm 0.09$	A
Decisión Tree	$91.79 \pm 0.11$	B
Ensemble	$91.94 \pm 0.11$	B
Redes Neuronales (RNN)	$96.89 \pm 0.14$	C
K - Nearest Neighbors (KNN)	$97.04 \pm 0.30$	C



**Figura 12.** Matriz de confusión, Precisión y Recall de los modelos: a) Decisión Tree, b) Support Vector Machine, c) K - Nearest Neighbors, d) Naive, e) Ensemble y f) Redes Neuronales.

### 4.3. Evaluación de la Precisión

Primero, el modelo de Árbol de Decisión. Este modelo tuvo una precisión del 90.7% para predecir a los que no pagan, 87.6% para predecir a los impuntuales, y 98.25%



para predecir a los puntuales. En este modelo, la mayor precisión se obtuvo al predecir a los puntuales.

En segundo lugar, el modelo de Máquina de Vectores de Soporte. Este modelo tuvo una precisión del 88.06% para predecir a los que no pagan, 85.94% para predecir a los impuntuales, y 99.09% para predecir a los puntuales. Al igual que en el modelo anterior, la mayor precisión se obtuvo al predecir a los puntuales.

El modelo de Vecinos más Cercanos (KNN) tuvo una precisión del 92.65% para predecir a los que no pagan, 99.15% para predecir a los impuntuales, y 100% para predecir a los puntuales. En este modelo, la mayor precisión se obtuvo al predecir a los puntuales.

El modelo de Bayes Ingenuo tuvo una precisión del 89.6% para predecir a los que no pagan, 84.44% para predecir a los impuntuales, y 100% para predecir a los puntuales. En este modelo, la mayor precisión se obtuvo al predecir a los puntuales.

El modelo de Conjunto tuvo una precisión del 86.43% para predecir a los que no pagan, 92.24% para predecir a los impuntuales, y 99.14% para predecir a los puntuales. En este modelo, la mayor precisión se obtuvo al predecir a los puntuales.

Finalmente, el modelo de Redes Neuronales tuvo una precisión del 93.94% para predecir a los que no pagan, 97.54% para predecir a los impuntuales, y 100% para predecir a los puntuales. Como en todos los demás modelos, la mayor precisión se obtuvo al predecir a los puntuales. La precisión responde a la pregunta ¿Qué proporción de la predicción de Pagadores Impuntuales o No Pagadores o Pagadores, con respecto al total de predicciones fue correcta?, es decir por ejemplo si uno de los modelos tiene una precisión de 0.9, es decir, cuando predice que un pagador es Impuntual, No Pagador o Pagador, es correcto el 90% de las veces de todas sus predicciones. La precisión fue superior a la de los estudios de (Baghdasaryan et al., 2022; Mesa et al., 2009) en los que se alcanzaron precisiones de 47.00 y 49.93% respectivamente, sin embargo, las exactitudes de estos estudios citados estuvieron muy cerca del 90%, esto, como se explicó anteriormente puede ser debido a la desproporción que hay entre las categorías, a diferencia de este estudio en el que las categorías de NO PAGADOR, IMPUNTUAL y PUNTUAL están equilibradas.

La selección del modelo estará en función a la categoría de pagador (No Paga, Impuntual o Paga) se desea predecir, por ejemplo si quisiera sensibilizar de manera personal a los contribuyentes que son IMPUNTUALES en sus impuestos para que paguen a tiempo, es posible que esté limitado en cuanto al número de personas que puedo visitar ya que podría estar limitado por personal, tiempo o combustible, por lo que necesitaría ser *preciso* en las predicciones que realizo, de acuerdo con la Tabla 4, debería utilizar el modelo K - Nearest Neighbors (KNN) cuya precisión es del 99.15%, es decir de cada 100 contribuyentes identificados como impuntuales, aproximadamente, solo 1 de ellos no sería un contribuyente impuntual, es decir que solo una visita sería fallida. De la misma manera con los contribuyentes que NO PAGAN, en este caso, se podría estar interesado en realizar visitas no solo de sensibilización, sino también de las sanciones que reciben las personas que no pagan sus tributos, en este caso utilizaría el modelo de Redes Neuronales (RNN) cuya precisión es del 93.94%, es decir de cada 100 visitas realizadas, aproximadamente solo 6 serían fallidas. El análisis detallado de la precisión de los modelos en las tres categorías —no pagadores, impuntuales y puntuales— proporciona una visión más profunda de su rendimiento. En todos los modelos, se observó la mayor precisión al predecir a los pagadores puntuales.

En el caso del modelo de Árbol de Decisión, se obtuvo una precisión del 98.25% para los pagadores puntuales, lo que concuerda con la investigación de Lessmann et al. (2015), donde se reconocía la capacidad de estos modelos para categorizar eficazmente los riesgos de crédito. El modelo de Máquina de Vectores de Soporte también mostró una precisión superior para los pagadores puntuales (99.09%), alineándose con los hallazgos de Huang et al. (2004) en cuanto a la eficacia de este modelo en la calificación crediticia. El modelo KNN mostró un rendimiento superior, con una precisión del 100% para los pagadores puntuales, superando los resultados reportados por Twala et al. (2008), donde KNN no superó a SVM y a las Redes Neuronales. Este resultado subraya la necesidad de más investigaciones para comprender mejor las circunstancias en las que KNN puede superar a otros modelos.

El modelo de Bayes Ingenuo también presentó una precisión del 100% para los pagadores puntuales, superando su desempeño en la predicción de los no pagadores e impuntuales. A pesar de que Louzada et al. (2016) encontraron que los modelos de Conjunto superan a los modelos Bayes Ingenuos en términos de exactitud, estos

resultados podrían sugerir que el modelo Bayes Ingenuo tiene su fortaleza en la identificación de pagadores puntuales. Para el modelo de Conjunto, aunque su precisión para los pagadores puntuales fue inferior a la de otros modelos (99.14%), aún superó la de los no pagadores e impuntuales. Finalmente, el modelo de Redes Neuronales demostró una precisión sobresaliente del 100% para los pagadores puntuales, alineándose con los hallazgos de Huang et al. (2004).

La Tabla 4 presenta los resultados de tres ANOVAs de Friedman separados, aplicados a las categorías de predicción 'No Pagadores', 'Impuntuales' y 'Puntuales' de algoritmos de aprendizaje automático, seguidos de pruebas post hoc de Nemenyi, cada uno con un nivel de significancia de  $\alpha = 0.05$ .

Para la categoría 'No Pagadores', el ANOVA de Friedman resultó en un valor de Chi-cuadrado ( $\chi^2$ ) de 48.46 con un p-valor menor que 0.01, lo que indica diferencias significativas en la precisión entre los algoritmos. La prueba post hoc de Nemenyi reveló que el algoritmo de Redes Neuronales (RNN) con un promedio de precisión de  $93.85 \pm 0.25$  está en el grupo estadístico F y muestra el rendimiento más alto, mientras que el algoritmo Ensemble con un promedio de  $86.27 \pm 0.17$  se encuentra en el grupo estadístico A, reflejando el rendimiento más bajo.

En el caso de los 'Impuntuales', el ANOVA de Friedman mostró un  $\chi^2$  de 48.97, con un p-valor también menor que 0.01, sugiriendo diferencias significativas en la precisión de los modelos. Aquí, el algoritmo K - Nearest Neighbors (KNN) con un promedio de  $99.10 \pm 0.12$  lidera en el grupo estadístico F, indicando la precisión más alta. Por otro lado, el algoritmo Naive Bayes con un promedio de  $84.35 \pm 0.19$  está en el grupo estadístico A, mostrando la precisión más baja.

Finalmente, para los 'Puntuales', se observó un  $\chi^2$  de 34.49 con un p-valor menor que 0.01 en el ANOVA de Friedman, lo que sugiere diferencias significativas entre los algoritmos. La prueba post hoc de Nemenyi posiciona al algoritmo Decision Tree con un promedio de  $98.28 \pm 0.14$  en el grupo estadístico A, representando el rendimiento más alto. En contraste, Naive Bayes se encuentra en el grupo estadístico C con un promedio de  $99.9 \pm 0.04$ , indicando el rendimiento más bajo en esta categoría.

**Tabla 4.** ANOVA de Friedman de la Precisión para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de  $\alpha = 0.05$ .

Fuente <sup>1</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	169.6	5	33.92	48.46	< 0.01
Error	5.4	45	0.12		
Total	175	59			
Algoritmos <sup>1</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Ensemble		86.27 $\pm$ 0.17	A		
Support Vector Machine (SVM)		88.07 $\pm$ 0.21		B	
Naive Bayes		89.79 $\pm$ 0.25			C
Decisión Tree		90.68 $\pm$ 0.20			D
K - Nearest Neighbors (KNN)		92.60 $\pm$ 0.20			E
Redes Neuronales (RNN)		93.85 $\pm$ 0.25			F
Fuente <sup>2</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	171.4	5	34.28	48.97	< 0.01
Error	3.6	45	0.08		
Total	175	59			
Algoritmos <sup>2</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Naive Bayes		84.35 $\pm$ 0.19	A		
Support Vector Machine (SVM)		85.95 $\pm$ 0.09		B	
Decisión Tree		87.00 $\pm$ 0.25			C
Ensemble		92.09 $\pm$ 0.22			D
Redes Neuronales (RNN)		97.90 $\pm$ 0.24			E
K - Nearest Neighbors (KNN)		99.10 $\pm$ 0.12			F
Fuente <sup>3</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	114.15	5	22.83	34.49	< 0.01
Error	51.35	45	1.14		
Total	165.5	59			
Algoritmos <sup>3</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Decisión Tree		98.28 $\pm$ 0.14	A		
Support Vector Machine (SVM)		99.00 $\pm$ 0.22			B
Ensemble		99.34 $\pm$ 0.12			B
Redes Neuronales (RNN)		99.88 $\pm$ 0.04			C
K - Nearest Neighbors (KNN)		99.80 $\pm$ 0.15			C
Naive Bayes		99.90 $\pm$ 0.04			C

<sup>1</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘No pagadores’

<sup>2</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Impuntuales’

<sup>3</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Pagadores’

#### 4.4. Evaluación del Recall

Primero, el modelo de Árbol de Decisión. Este modelo tuvo un Recall del 92.13% para los clasificados como No Paga, 88.98% para los clasificados como Impuntual, y 94.92% para los clasificados como Puntual. En este modelo, el mayor Recall se obtuvo con los puntuales.

Luego, el modelo de Máquina de Vectores de Soporte. Este modelo tuvo un Recall del 92.91% para los clasificados como No Paga, 86.61% para los clasificados como Impuntual, y 92.37% para los clasificados como Puntual. A diferencia del modelo anterior, el mayor Recall se obtuvo con los que no pagan.

El modelo de Vecinos más Cercanos (KNN) tuvo un Recall del 99.21% para los clasificados como No Paga, 92.13% para los clasificados como Impuntual, y 100% para los clasificados como Puntual. En este modelo, el mayor Recall se obtuvo con los puntuales. El modelo de Bayes Ingenuo tuvo un Recall del 88.19% para los clasificados como No Paga, 89.76% para los clasificados como Impuntual, y 94.92% para los clasificados como Puntual. En este modelo, el mayor Recall se obtuvo con los puntuales.

El modelo de Conjunto tuvo un Recall del 95.28% para los clasificados como No Paga, 84.25% para los clasificados como Impuntual, y 97.46% para los clasificados como Puntual. En este modelo, el mayor Recall se obtuvo con los puntuales. Finalmente, el modelo de Redes Neuronales tuvo un Recall del 97.64% para los clasificados como No Paga, 93.7% para los clasificados como Impuntual, y 100% para.

La sensibilidad o Recall en los estudios de (Baghdasaryan et al., 2022; López et al., 2019; Mesa et al., 2009) la sensibilidad fue de 86.4%, 98.60% y 82.21%, para los modelos de Decisión Tree, Regresión Logística y Redes Neuronales respectivamente, estas sensibilidades son similares a la de esta tesis, para Decisión Tree se encontraron sensibilidades desde 88.98% hasta 94.92% siendo ligeramente superiores, sin embargo para el modelo de redes neuronales se encontraron sensibilidades desde 93.7% hasta 100%, siendo este rango notablemente mayor al del estudio de (Mesa et al., 2009).

La sensibilidad responde a la pregunta ¿qué porcentaje de los que realmente son no pagadores/pagadores o impuntuales se puede identificar dada una muestra?, esta pregunta puede ser de interés si se quiere hacer proyecciones del grado de recaudación que se tendrá para un futuro, en este caso resulta interesante identificar a los que son

PUNTUALES en primer lugar para el análisis del grado de recaudación temprana, en estos casos los modelos que mejor sensibilidad tienen fueron los de K - Nearest Neighbors y Redes Neuronales en los que se podría identificar al 100% de los pagadores puntuales en una muestra. Puede ser que también se esté interesado en la recaudación tardía, es decir los ingresos fiscales tardíos, en este caso es de interés identificar a los IMPUNTUALES, y en este caso el modelo recomendado sería el de Redes Neuronales cuya sensibilidad o Recall es del 93.7%, es decir que de cada 100 contribuyentes IMPUNTUALES, se lograrán identificar correctamente aproximadamente 94.

La Tabla 5 detalla los resultados de tres ANOVAs de Friedman distintos, efectuados para evaluar la sensibilidad (Recall) de algoritmos de aprendizaje automático en las categorías de 'No Pagadores', 'Impuntuales' y 'Pagadores', seguidos por pruebas post hoc de Nemenyi, aplicando un nivel de significancia de  $\alpha = 0.05$ . En la categoría 'No Pagadores', se obtuvo un valor Chi-cuadrado de 49.09 y un p-valor menor que 0.01, lo cual señala diferencias significativas en la sensibilidad entre los algoritmos. La prueba post hoc de Nemenyi indicó que el algoritmo K - Nearest Neighbors (KNN) con un promedio de sensibilidad de  $99.13 \pm 0.15$  se sitúa en el grupo estadístico E, mostrando el nivel más alto de sensibilidad. En contraste, Naive Bayes, con un promedio de  $88.11 \pm 0.14$ , se ubica en el grupo estadístico A, reflejando la sensibilidad más baja. Respecto a los 'Impuntuales', el análisis arrojó un valor Chi-cuadrado de 49.49 con un p-valor también inferior a 0.01, sugiriendo diferencias significativas. En este caso, las Redes Neuronales (RNN) con una sensibilidad de  $93.81 \pm 0.14$  ocupan el grupo estadístico F, representando la mayor sensibilidad detectada, mientras que Ensemble se encuentra en el grupo estadístico A con una sensibilidad de  $84.17 \pm 0.12$ , siendo la más baja de la categoría. Para los 'Pagadores', el ANOVA de Friedman presentó un valor Chi-cuadrado de 46.39 y un p-valor menor que 0.01. La prueba post hoc de Nemenyi revela que tanto K - Nearest Neighbors (KNN) como Redes Neuronales (RNN) comparten el grupo estadístico E con sensibilidades de  $99.71 \pm 0.13$  y  $99.72 \pm 0.04$  respectivamente, mostrando las sensibilidades más altas. Por otro lado, Support Vector Machine (SVM) con un promedio de  $91.77 \pm 0.26$  se posiciona en el grupo estadístico A, indicando la sensibilidad más baja dentro de esta categoría.

**Tabla 5.** ANOVA de Friedman de la Sensibilidad (Recall) para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de  $\alpha = 0.05$ .

Fuente <sup>1</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	171.8	5	34.36	49.09	< 0.01
Error	3.2	45	0.07		
Total	175	59			
Algoritmos <sup>1</sup>	Promedio $\pm$ E.E.		Grupos estadísticos		
Naive Bayes	88.11 $\pm$ 0.14		A		
Decisión Tree	92.45 $\pm$ 0.16		B		
Support Vector Machine (SVM)	92.98 $\pm$ 0.05		B		
Ensemble	94.97 $\pm$ 0.15		C		
Redes Neuronales (RNN)	97.55 $\pm$ 0.18		D		
K - Nearest Neighbors (KNN)	99.13 $\pm$ 0.15		E		
Fuente <sup>2</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	173.2	5	34.64	49.49	< 0.01
Error	1.8	45	0.04		
Total	175	59			
Algoritmos <sup>2</sup>	Promedio $\pm$ E.E.		Grupos estadísticos		
Ensemble	84.17 $\pm$ 0.12		A		
Support Vector Machine (SVM)	86.65 $\pm$ 0.17		B		
Decisión Tree	89.28 $\pm$ 0.21		C		
Naive Bayes	89.79 $\pm$ 0.09		D		
K - Nearest Neighbors (KNN)	92.17 $\pm$ 0.12		E		
Redes Neuronales (RNN)	93.81 $\pm$ 0.14		F		
Fuente <sup>3</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	161.45	5	32.29	46.39	< 0.01
Error	12.55	45	0.28		
Total	174	59			
Algoritmos <sup>3</sup>	Promedio $\pm$ E.E.		Grupos estadísticos		
Support Vector Machine (SVM)	91.77 $\pm$ 0.26		A		
Decisión Tree	94.89 $\pm$ 0.11		B		
Naive Bayes	94.87 $\pm$ 0.24		C		
Ensemble	97.67 $\pm$ 0.41		D		
K - Nearest Neighbors (KNN)	99.71 $\pm$ 0.13		E		
Redes Neuronales (RNN)	99.92 $\pm$ 0.04		E		

<sup>1</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘No pagadores’

<sup>2</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Impuntuales’

<sup>3</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Pagadores’

#### 4.5. Puntuación F1 o Desempeño

Primero, el modelo de Árbol de Decisión. Este modelo tuvo una Puntuación F1 del 91.41% para los clasificados como No Paga, 88.28% para los clasificados como Impuntual, y 96.56% para los clasificados como Puntual, con un promedio de 92.08%. En este modelo, la mayor Puntuación F1 se obtuvo con los puntuales.

Luego, el modelo de Máquina de Vectores de Soporte. Este modelo tuvo una Puntuación F1 del 90.42% para los clasificados como No Paga, 86.27% para los clasificados como Impuntual, y 95.61% para los clasificados como Puntual, con un promedio de 90.77%. A diferencia del modelo anterior, la mayor Puntuación F1 se obtuvo con los que no pagan.

El modelo de Vecinos más Cercanos (KNN) tuvo una Puntuación F1 del 95.82% para los clasificados como No Paga, 95.51% para los clasificados como Impuntual, y 100% para los clasificados como Puntual, con un promedio de 97.11%. En este modelo, la mayor Puntuación F1 se obtuvo con los puntuales.

El modelo de Bayes Ingenuo tuvo una Puntuación F1 del 88.89% para los clasificados como No Paga, 87.02% para los clasificados como Impuntual, y 97.39% para los clasificados como Puntual, con un promedio de 91.10%. En este modelo, la mayor Puntuación F1 se obtuvo con los puntuales.

El modelo de Conjunto tuvo una Puntuación F1 del 90.64% para los clasificados como No Paga, 88.06% para los clasificados como Impuntual, y 98.29% para los clasificados como Puntual, con un promedio de 92.33%. En este modelo, la mayor Puntuación F1 se obtuvo con los puntuales.

Finalmente, el modelo de Redes Neuronales tuvo una Puntuación F1 del 95.75% para los clasificados como No Paga, 95.58% para los clasificados como Impuntual, y 100% para los clasificados como Puntual, con un promedio de 97.11%.

La Tabla 6 presenta los resultados de tres ANOVAs de Friedman aplicados a la puntuación F1 o desempeño de varios algoritmos de aprendizaje automático para las categorías de 'No Pagadores', 'Impuntuales' y 'Pagadores', seguidos por pruebas post hoc de Nemenyi con un nivel de significancia de  $\alpha = 0.05$ .



**Tabla 6.** ANOVA de Friedman de la Puntuación F1 o Desempeño para los ‘No Pagadores’, ‘Impuntuales’ y ‘Pagadores’ respectivamente, de los Algoritmos de Machine Learning y prueba post hoc de Nemenyi, ambas a un nivel de significancia de  $\alpha = 0.05$ .

Fuente <sup>1</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	160	5	32	45.71	< 0.01
Error	15	45	0.33		
Total	175	59			
Algoritmos <sup>1</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Naive Bayes		88.67 $\pm$ 0.13	A		
Support Vector Machine (SVM)		90.43 $\pm$ 0.17	B		
Ensemble		90.68 $\pm$ 0.25	B		
Decisión Tree		91.5 $\pm$ 0.08	C		
K - Nearest Neighbors (KNN)		95.56 $\pm$ 0.21	D		
Redes Neuronales (RNN)		95.55 $\pm$ 0.31	D		
Fuente <sup>2</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	161.2	5	32.24	46.06	< 0.01
Error	13.8	45	0.31		
Total	175	59			
Algoritmos <sup>2</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Support Vector Machine (SVM)		86.15 $\pm$ 0.2	A		
Naive Bayes		87 $\pm$ 0.1	A		
Ensemble		87.75 $\pm$ 0.16	B		
Decisión Tree		88.07 $\pm$ 0.15	B		
Redes Neuronales (RNN)		95.49 $\pm$ 0.17	C		
K - Nearest Neighbors (KNN)		95.67 $\pm$ 0.05	C		
Fuente <sup>3</sup>	SC	Gl	SM	Chi-sq	p-valor
Algoritmos	154.75	5	30.95	44.34	< 0.01
Error	19.75	45	0.44		
Total	174.5	59			
Algoritmos <sup>3</sup>		Promedio $\pm$ E.E.	Grupos estadísticos		
Support Vector Machine (SVM)		95.73 $\pm$ 0.13	A		
Decisión Tree		96.55 $\pm$ 0.13	B		
Naive Bayes		96.77 $\pm$ 0.18	B		
Ensemble		97.98 $\pm$ 0.29	C		
K - Nearest Neighbors (KNN)		99.83 $\pm$ 0.07	D		
Redes Neuronales (RNN)		99.73 $\pm$ 0.15	E		

<sup>1</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘No pagadores’

<sup>2</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Impuntuales’

<sup>3</sup> ANOVA de Friedman y prueba post hoc de Nemenyi para los ‘Pagadores’

En la categoría 'No Pagadores', se observa un valor Chi-cuadrado de 45.71 y un p-valor menor que 0.01, lo que indica diferencias significativas en la puntuación F1 entre

los algoritmos. La prueba post hoc de Nemenyi coloca a los algoritmos K - Nearest Neighbors (KNN) y Redes Neuronales (RNN) ambos en el grupo estadístico D, con puntuaciones F1 de  $95.56 \pm 0.21$  y  $95.55 \pm 0.31$  respectivamente, indicando los niveles de desempeño más altos. Por otro lado, Naive Bayes, con una puntuación F1 de  $88.67 \pm 0.13$ , pertenece al grupo estadístico A, reflejando el menor desempeño en esta categoría.

Para los 'Impuntuales', el ANOVA de Friedman muestra un valor Chi-cuadrado de 46.06 y un p-valor menor que 0.01. Aquí, K - Nearest Neighbors (KNN) presenta la puntuación F1 más alta con  $95.67 \pm 0.05$ , clasificado en el grupo estadístico C, mientras que Support Vector Machine (SVM) y Naive Bayes están en el grupo estadístico A con puntuaciones F1 de  $86.15 \pm 0.2$  y  $87 \pm 0.1$ , mostrando los niveles más bajos de desempeño.

Finalmente, en la categoría 'Pagadores', el ANOVA de Friedman resulta en un valor Chi-cuadrado de 44.34 y un p-valor menor que 0.01. La prueba post hoc de Nemenyi identifica al algoritmo K - Nearest Neighbors (KNN) con una puntuación F1 de  $99.83 \pm 0.07$  en el grupo estadístico D, ostentando el mayor desempeño. Support Vector Machine (SVM), aunque con una puntuación F1 alta de  $95.73 \pm 0.13$ , se encuentra en el grupo estadístico A, indicando el menor desempeño en comparación con el resto de los algoritmos en esta categoría.

## V. CONCLUSIÓN

Se concluye que existen diferencias estadísticamente significativas (ANOVA de Friedman p-valor < 0.01) en el desempeño o Puntuación F1 de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019. Los modelos que tuvieron mejor desempeño en la categoría de 'No Pagadores', 'Impuntuales' y 'Pagadores' fueron K - Nearest Neighbors, K - Nearest Neighbors y Redes Neuronales, respectivamente.

- Los desempeños de los 6 modelos de machine learning se obtuvieron a partir de una base de datos en donde se recogió información socioeconómica y las categorías de pagador de impuesto predial de 372 contribuyentes, en el entrenamiento los modelos recibieron validación por partición k – fold de 5, se calculó el índice F1 como métrica de desempeño.
- Estos resultados tienen el potencial de influir en la política fiscal y en las estrategias de cobranza, al proporcionar una herramienta efectiva para identificar y categorizar a los contribuyentes. Esto podría conducir a una mayor eficiencia en la recaudación de impuestos y a una mejor comprensión de los patrones de pago dentro de la comunidad. En última instancia, la aplicación de estos modelos podría traducirse en una mejor planificación y asignación de recursos, beneficiando tanto al gobierno local como a los ciudadanos de Tingo María.

Se concluye que existen diferencias estadísticamente significativas (ANOVA de Friedman p-valor < 0.01) en la Exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019. El modelo que tuvo mejor desempeño para esta métrica fue K - Nearest Neighbors.

- Los desempeños de los 6 modelos de machine learning se obtuvieron a partir de una base de datos en donde se recogió información socioeconómica y las categorías de pagador de impuesto predial de 372 contribuyentes, en el entrenamiento los modelos recibieron validación por partición k – fold de 5, se calculó la exactitud como métrica de desempeño.
- Los resultados de los desempeños de los modelos con respecto a la métrica de exactitud revelan un rango notablemente alto, con K-Nearest Neighbors y Redes Neuronales destacando con exactitudes de 97.04% y 97.01% respectivamente.

Se concluye que existen diferencias estadísticamente significativas (ANOVA de Friedman  $p$ -valor  $< 0.01$ ) en la Precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019. Los modelos que tuvieron mejor desempeño en la categoría de 'No Pagadores', 'Impuntuales' y 'Pagadores' fueron Redes Neuronales, K - Nearest Neighbors y Naive Bayes, respectivamente.

- Los desempeños de los 6 modelos de machine learning se obtuvieron a partir de una base de datos en donde se recogió información socioeconómica y las categorías de pagador de impuesto predial de 372 contribuyentes, en el entrenamiento los modelos recibieron validación por partición k – fold de 5, se calculó la precisión como métrica de desempeño.
- Para el uso de los modelos con respecto a la sensibilidad, la selección del modelo de machine learning se hará en función de la categoría de pagador que se desee predecir (No Paga, Impuntual o Paga). Si se quiere enfocar en los contribuyentes impuntuales en sus impuestos, se enfrentará a limitaciones como personal, tiempo o combustible. Por lo tanto, se necesita una precisión alta en las predicciones para optimizar los recursos. En este contexto, el modelo K-Nearest Neighbors (KNN) sería la elección óptima, ya que su precisión es del 99.15%. Esto significa que de cada 100 contribuyentes identificados como impuntuales, solo 1 de ellos sería una identificación incorrecta, lo que haría que solo una visita fuera fallida.

Se concluye que existen diferencias estadísticamente significativas (ANOVA de Friedman  $p$ -valor  $< 0.01$ ) en la Sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019. Los modelos que tuvieron mejor desempeño en la categoría de 'No Pagadores', 'Impuntuales' y 'Pagadores' fueron Redes Neuronales, Redes Neuronales y K - Nearest Neighbors, respectivamente.

- Los desempeños de los 6 modelos de machine learning se obtuvieron a partir de una base de datos en donde se recogió información socioeconómica y las categorías de pagador de impuesto predial de 372 contribuyentes, en el entrenamiento los modelos recibieron validación por partición k – fold de 5, se calculó la sensibilidad como métrica de desempeño.

- La sensibilidad se refiere a la capacidad de identificar un porcentaje específico de una categoría, como no pagadores, pagadores o impuntuales, en una muestra. Esto es relevante para las proyecciones de recaudación futura, particularmente en la identificación de pagadores puntuales. En este contexto, los modelos de K-Nearest Neighbors y Redes Neuronales se destacan, ya que pueden identificar al 100% de pagadores puntuales en una muestra.

## VI. REFERENCIAS BIBLIOGRÁFICAS

- Aceituno Rojo, M. R. (2019). *Modelo predictivo de análisis de riesgo crediticio usando Machine Learning en una entidad del sector microfinanciero* [Tesis doctoral]. Universidad Nacional del Altiplano.
- Alvarado Baena, L. (2023). *El efecto de la corrupción sobre la recaudación tributaria*. Ruiz Healy Times. <https://ruizhealytimes.com/liliana/el-efecto-de-la-corrupcion-sobre-la-recaudacion-tributaria/>
- Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer's Network Data in Fraud Detection. *Applied Artificial Intelligenc*, 36(1). <https://doi.org/10.1080/08839514.2021.2012002>
- BCRP. (2021). *Reporte de Inflación Marzo 2019*. <https://www.bcrp.gob.pe/publicaciones/reportes-de-inflacion/reportes-de-inflacion-marzo-2019.html>
- Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing*, 19(1), 168–184. <https://doi.org/10.2308/aud.2000.19.1.169>
- Cayatopa Rivera, L. E. (2021). Perú: Determinación de la brecha de recaudación potencial de los impuestos y arbitrios municipales, 2014-2021. *Quipukamayoc*, 29(60), 61–72. <https://doi.org/10.15381/QUIPU.V29I60.20535>
- Cerullo, M. J., & Cerullo, V. (1999). Using neural networks to predict financial reporting fraud: Part 1. *Computer Fraud and Security*, 1999(5), 14–17. [https://doi.org/10.1016/s1361-3723\(99\)80015-3](https://doi.org/10.1016/s1361-3723(99)80015-3)
- Chávez Alvarez, M. (2021). *Cultura tributaria, factores socioeconomicos y su influencia en la recaudación del impuesto predial en el distrito de Jesús, año 2020* [Tesis para Contador Público]. Universidad Privada del Norte.
- Çiğşar, B., & Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming*, 2019. <https://doi.org/10.1155/2019/8706505>

- Didimo, W., Grilli, L., Liotta, G., Menconi, L., Montecchiani, F., & Pagliuca, D. (2020). Combining Network Visualization and Data Mining for Tax Risk assessment. *IEEE Access*, 8, 16073–16086. <https://doi.org/10.1109/ACCESS.2020.2967974>
- Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, 27(3), 3743–3775. <https://doi.org/10.1007/S10639-021-10741-7/METRICS>
- FMI. (2023). *Los impuestos pueden apoyar el crecimiento y reducir la desigualdad en América Latina y el Caribe*. <https://www.imf.org/es/News/Articles/2021/12/10/na121021-Taxes-Support-Growth-Reduce-Inequality-Latin-America-Caribbean>
- Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1–2), 107–143. <https://doi.org/10.1023/A:1008699112516/METRICS>
- Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. *Proceedings of the Hawaii International Conference on System Sciences*, 3, 621–630. <https://doi.org/10.1109/hicss.1994.323314>
- Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, 100577. <https://doi.org/10.1016/J.JBEF.2021.100577>
- Han, J., Kamber, M., Pei, J., & Kaufmann, M. (2011). *Data Mining: Concepts and Techniques* (Third).
- Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4), 543–558. [https://doi.org/10.1016/S0167-9236\(03\)00086-1](https://doi.org/10.1016/S0167-9236(03)00086-1)

- Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4), 995–1003. <https://doi.org/10.1016/j.eswa.2006.02.016>
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/J.EJOR.2015.05.030>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors (Basel, Switzerland)*, 18(8). <https://doi.org/10.3390/S18082674>
- López, C. P., Rodríguez, M. J. D., & Santos, S. de L. (2019). Tax fraud detection through neural networks: An application using a sample of personal income taxpayers. *Future Internet*, 11(4), 86. <https://doi.org/10.3390/FI11040086>
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117–134. <https://doi.org/10.1016/J.SORMS.2016.10.001>
- McKinsey. (2023). *Intelligent process automation: The engine at the core of the next-generation operating model* / McKinsey. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/intelligent-process-automation-the-engine-at-the-core-of-the-next-generation-operating-model>
- MEF, SECO, & GIZ. (2016). *Guía para el registro y determinación del impuesto predial*.
- Mesa, F. R., Raineri, A., Maturana, S., & Kaempffer, A. M. (2009). Fraudes a los sistemas de salud en Chile: Un modelo para su detección. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health*, 25(1), 56–61. <https://doi.org/10.1590/s1020-49892009000100009>
- Meza, K. (2019). ¿Qué es el impuesto predial y cómo se paga? In *EL COMERCIO. NOTICIAS EL COMERCIO PERÚ*.
- Montero Flores, J. A. (2022). *Influencia de los Factores Socioeconómicos en el Cumplimiento del Pago del Impuesto Predial en Contribuyentes de la Municipalidad*



- Provincial de Sullana, Año 2020* [Tesis de Ingeniero Economista, Universidad Nacional de Frontera]. <http://repositorio.unf.edu.pe/handle/UNF/145>
- Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569. <https://doi.org/10.1016/j.dss.2010.08.006>
- OCDE. (2020). Estadísticas tributarias en América Latina y el Caribe 2020 - Ecuador La recaudación tributaria como porcentaje del PIB La recaudación tributaria como porcentaje del PIB en comparación con otros países de América Latina y el Rango Países de América Latina. In *Organización para la Cooperación y el Desarrollo Económico*. <http://www.oecd.org/tax/tax-policy/estadisticas-tributarias-america-latina-caribe-mexico.pdf>
- Perbendaharaan, J., Negara Dan Kebijakan Publik, K., David Febriminanto, R., & Wasesa, M. (2022). Machine Learning Analytics for Predicting Tax Revenue Potential. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara Dan Kebijakan Publik*, 7(3), 193–205. <https://doi.org/10.33105/ITREV.V7I3.497>
- Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Computers in Human Behavior*, 28(3), 1002–1013. <https://doi.org/10.1016/j.chb.2012.01.002>
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. [https://doi.org/10.1056/NEJMRA1814259/SUPPL\\_FILE/NEJMRA1814259\\_DISCLOSURES.PDF](https://doi.org/10.1056/NEJMRA1814259/SUPPL_FILE/NEJMRA1814259_DISCLOSURES.PDF)
- Robert, C. (2014). Machine Learning, a Probabilistic Perspective. <https://doi.org/10.1080/09332480.2014.914768>, 27(2), 62–63. <https://doi.org/10.1080/09332480.2014.914768>
- Romero Carazas, R., Soria Diaz, M. E., Del Castillo Gómez, W., & Colmenares de Zavala, Y. (2021). El Estudio del impuesto predial en gobiernos locales del Perú. *ECA Sinergia*, 12(3), 124. [https://doi.org/10.33936/ECA\\_SINERGIA.V12I3.3583](https://doi.org/10.33936/ECA_SINERGIA.V12I3.3583)

- Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, 148, 45–54. <https://doi.org/10.1016/j.procs.2019.01.007>
- Sharma, A., & Panigrahi, P. K. (2013). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications*, 39(1), 37–47. <https://doi.org/10.5120/4787-7016>
- Smolka, M., & De Cesare, C. M. (2012). Property Tax and Informal Property. *A Primer on Property Tax: Administration and Policy*, 265–286. <https://doi.org/10.1002/9781118454343.CH12>
- Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, 17(4), 179–191. <https://doi.org/10.1108/02686900210424321>
- Tong, L. (2019). *Análisis Comparativo de Técnicas de Aprendizaje Automático para Detectar Fraude en Tarjetas de Crédito*. Universidad Católica San Pablo.
- Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, 1(1), 100002. <https://doi.org/10.1016/J.JJIMEI.2020.100002>
- Yeh, I. C., & hui Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2 PART 1), 2473–2480. <https://doi.org/10.1016/j.eswa.2007.12.020>

## ANEXOS

## ANEXO 1



**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**  
**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**  
**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA**  
**Y SISTEMAS**



Buen día estimado(a) señor(a) cliente de contribuyente, a continuación, se le hará una encuesta en la que se busca conocer más de los contribuyentes, para ello se le formularán un total de 9 preguntas. El tratamiento de los datos que aquí figuren será confidencial y únicamente con fines de investigación, por lo que no se facilitarán a ninguna persona u organización ajena a la misma. La participación es totalmente con fines de investigación, por lo que le rogamos la máxima sinceridad. Gracias.

### CUESTIONARIO

1. ¿Cuál es el ingreso familiar mensual (S/.)?
  - <1050
  - 1050 - 2500
  - 2500 - 5000
  - 5000
  
2. ¿Considera usted que su situación económica impide el pago del impuesto predial?
  - Siempre
  - Casi siempre
  - A veces
  - Nunca
  
3. ¿Usted percibe que las autoridades de la municipalidad cumplen con solucionar los problemas de la provincia?
  - Totalmente en desacuerdo
  - En desacuerdo
  - Neutral
  - De acuerdo
  - Totalmente de acuerdo

4. ¿Usted cree que los funcionarios de la Municipalidad están inmersos en actos de corrupción?
- Totalmente en desacuerdo
  - En desacuerdo
  - Neutral
  - De acuerdo
  - Totalmente de acuerdo
5. ¿Acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente?
- SI
  - NO
6. ¿Es consciente de las sanciones o multas que puede generar el no declarar información de sus predios?
- SI
  - NO
7. ¿Considera usted que es primordial el pago del impuesto predial?
- SI
  - NO
8. ¿Tiene usted una estabilidad laboral mayor a 3 años?
- SI
  - NO
9. ¿Considera usted que se brindan servicios públicos en la zona en la que vive?
- SI
  - NO

## ANEXO 2. MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	DISEÑO	MUESTRA
<b>GENERAL</b>			<b>Independiente:</b> Desempeño de modelos de Machine Learning: <b>Dimensiones:</b> Exactitud Precisión Sensibilidad (Recall)	La presente investigación se clasifica como una investigación aplicada, dado que se orienta hacia la resolución de un problema práctico específico: predicción de pagadores del impuesto predial en la ciudad de Tingo María. La investigación aplicada se enfoca en la utilización de teorías, principios y técnicas con el objetivo de abordar y resolver problemas concretos del mundo real. En este contexto, se implementan algoritmos de Machine Learning para identificar patrones de fraude.	En la ciudad de Tingo María existen 12821 contribuyentes y de acuerdo con la base de datos de La Municipalidad Provincial de Leoncio Prado, 4365 pagan el impuesto predial de manera puntual, 187 pagan con retrasos y 8269 no pagan.
¿Existe diferencia en el desempeño de modelos en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019?	Determinar si existe la diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019	Existe diferencia en el desempeño de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019, es significativa			
<b>ESPECÍFICOS</b>					
¿Existe diferencia en la exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019?	Evaluar la diferencia en la exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019	Existe diferencia en la exactitud de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019, es significativa	<b>Dependiente:</b> Modelos de Machine Learning <b>Dimensiones:</b> Decisión Tree Support Vector Machine K - Nearest Neighbors Naive Bayes Ensemble Redes Neuronales		
¿Existe diferencia en la precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019?	Determinar si existe la diferencia en la precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019	Existe diferencia en la precisión de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019, es significativa			
¿Existe diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019?	Determinar si existe la diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019	Existe diferencia en la sensibilidad de modelos de Machine Learning en predicción de pagadores del impuesto predial en la ciudad de Tingo María para el año 2019, es significativa			



**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**  
**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**  
**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA**  
**Y SISTEMAS**



**ANEXO 3: VALIDACIÓN DE EXPERTOS**

**Título:** Evaluación comparativa en el desempeño de modelos de machine learning en la predicción de pagadores del impuesto predial de la ciudad de Tingo María para el año 2019

Nombre de Juez	<i>García Villegas Christian</i>
Institución donde labora	<i>UNAS - FUS</i>
Cargo	<i>Docente</i>

Tingo María, Huánuco, 2023

**Autor:** Antonio Gabriel Junco Aranciaga

**Instrumento motivo de evaluación:** Cuestionario

**Detalles del instrumento:** El instrumento se ha elaborado para conocer el perfil socioeconómico del tipo de pagador de impuesto predial y a partir de eso predecir si la persona pagará, no pagará o pagará tarde su impuesto predial.

Calificar con:

- (1) No cumple con el criterio
- (2) Bajo nivel
- (3) Moderado nivel
- (4) Alto nivel

Recomendaciones

**DECISIÓN DEL EXPERTO:**

Aplicable

Aplicable después de corregir

No aplicable



*Christian García Villegas*  
 Mg. Christian García Villegas  
 DOCENTE

Firma del experto

Perfil Socioeconómico del pagador de impuesto	Relevancia	Coherencia	Suficiencia	Claridad
¿Cuál es el ingreso familiar mensual (S/.)?	4	4	4	4
¿Considera usted que su situación económica impide el pago del impuesto predial?	4	4	4	4
¿Usted percibe que las autoridades de la municipalidad cumplen con solucionar los problemas de la provincia?	3	2	2	2
¿Usted cree que los funcionarios de la Municipalidad están inmersos en actos de corrupción?	2	2	2	2
¿Acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente?	4	4	4	4
¿Es consciente de las sanciones o multas que puede generar el no declarar información de sus predios?	3	3	3	3
¿Considera usted que es primordial el pago del impuesto predial?	4	4	4	4
¿Tiene usted una estabilidad laboral mayor a 3 años?	3	2	3	2
¿Considera usted que se brindan servicios públicos en la zona en la que vive?	3	2	2	2
	3	3	3	3



**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**  
**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**  
**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA**  
**Y SISTEMAS**



**ANEXO 3: VALIDACIÓN DE EXPERTOS**

**Título:** Evaluación comparativa en el desempeño de modelos de machine learning en la predicción de pagadores del impuesto predial de la ciudad de Tingo María para el año 2019

Nombre de Juez	Rannoverng Janac Mandesino
Institución donde labora	UNAS
Cargo	Docente

Tingo María, Huánuco, 2023

**Autor:** Antonio Gabriel Junco Aranciaga

**Instrumento motivo de evaluación:** Cuestionario

**Detalles del instrumento:** El instrumento se ha elaborado para conocer el perfil socioeconómico del tipo de pagador de impuesto predial y a partir de eso predecir si la persona pagará, no pagará o pagará tarde su impuesto predial.

Calificar con:

- (1) No cumple con el criterio
- (2) Bajo nivel
- (3) Moderado nivel ✓
- (4) Alto nivel

**Recomendaciones** Centrarse en mas datos para impuesto predial

**DECISIÓN DEL EXPERTO:**

Aplicable

Aplicable después de corregir

No aplicable

Firma del experto



Perfil Socioeconómico del pagador de impuesto	Relevancia	Coherencia	Suficiencia	Claridad
¿Cuál es el ingreso familiar mensual (S/.)?	4	4	4	4
¿Considera usted que su situación económica impide el pago del impuesto predial?	4	4	4	4
¿Usted percibe que las autoridades de la municipalidad cumplen con solucionar los problemas de la provincia?	2	2	2	3
¿Usted cree que los funcionarios de la Municipalidad están inmersos en actos de corrupción?	2	2	2	3
¿Acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente?	4	4	4	4
¿Es consciente de las sanciones o multas que puede generar el no declarar información de sus predios?	2	2	2	3
¿Considera usted que es primordial el pago del impuesto predial?	4	4	4	4
¿Tiene usted una estabilidad laboral mayor a 3 años?	3	3	3	3
¿Considera usted que se brindan servicios públicos en la zona en la que vive?	2	2	2	2
$\bar{M}$	3	3	3	3



**UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA**  
**FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**  
**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA**  
**Y SISTEMAS**



**ANEXO 3: VALIDACIÓN DE EXPERTOS**

**Título:** Evaluación comparativa en el desempeño de modelos de machine learning en la predicción de pagadores del impuesto predial de la ciudad de Tingo María para el año 2019

Nombre de Juez	José Orlando Castillo Cornelio
Institución donde labora	UNAS
Cargo	Docente

Tingo María, Huánuco, 2023

**Autor:** Antonio Gabriel Junco Aranciaga

**Instrumento motivo de evaluación:** Cuestionario

**Detalles del instrumento:** El instrumento se ha elaborado para conocer el perfil socioeconómico del tipo de pagador de impuesto predial y a partir de eso predecir si la persona pagará, no pagará o pagará tarde su impuesto predial.

Calificar con:

- (1) No cumple con el criterio
- (2) Bajo nivel
- (3) Moderado nivel ✓
- (4) Alto nivel

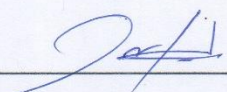
**Recomendaciones:** En lo posible agregar datos que permitan medir la disposición a pagar el impuesto predial.

**DESICIÓN DEL EXPERTO:**

( ) Aplicable

(x) Aplicable después de corregir

( ) No aplicable

  
 Firma del experto

Perfil Socioeconómico del pagador de impuesto	Relevancia	Coherencia	Suficiencia	Claridad
¿Cuál es el ingreso familiar mensual (S/.)?	4	4	4	4
¿Considera usted que su situación económica impide el pago del impuesto predial?	4	4	4	4
¿Usted percibe que las autoridades de la municipalidad cumplen con solucionar los problemas de la provincia?	2	2	2	2
¿Usted cree que los funcionarios de la Municipalidad están inmersos en actos de corrupción?	3	3	2	2
¿Acepta la tributación como un compromiso que forma parte de sus deberes y derechos como contribuyente?	4	4	4	4
¿Es consciente de las sanciones o multas que puede generar el no declarar información de sus predios?	3	2	2	2
¿Considera usted que es primordial el pago del impuesto predial?	4	4	4	4
¿Tiene usted una estabilidad laboral mayor a 3 años?	3	3	2	2
¿Considera usted que se brindan servicios públicos en la zona en la que vive?	1	1	2	2
PT:	3	3	3	3