UNIVERSIDAD NACIONAL
"PEDRO RUIZ GALLO"

ESCUELA DE POSTGRADO

MAESTRÍA EN INGENIERIA DE SISTEMAS

OPTIMIZACIÓN DE LA CAMPAÑA DE AMNISTIA TRIBUTARIA DEL CENTRO DE GESTIÓN TRIBUTARIA DE CHICLAYO APLICANDO UN MODELO DE PREDICCIÓN BASADO EN EL COMPORTAMIENTO DE PAGO DE LOS CONTRIBUYENTES

TESIS

PRESENTADA PARA OPTAR EL GRADO ACADEMICO DE MAESTRO EN INGENIERIA DE SISTEMAS CON MENCION EN GERENCIA DE TECNOLOGIAS DE LA INFORMACION Y GESTION DEL SOFTWARE

AUTOR:

Ing. CARLOS ALBERTO INOÑAN GONZALES

ASESORA:

Ing. EMMA VIRGINIA NOBLECILLA MONTEALEGRE

LAMBAYEQUE – PERÚ

2018

OPTIMIZACIÓN DE LA CAMPAÑA DE AMNISTIA TRIBUTARIA DEL CENTRO DE GESTIÓN TRIBUTARIA DE CHICLAYO APLICANDO UN MODELO DE PREDICCIÓN BASADO EN EL COMPORTAMIENTO DE PAGO DE LOS CONTRIBUYENTES

ING. CARLOS A. INOÑÁN GONZÁLES
AUTOR

DRA. EMMA NOBLECILLA MONTEALEGRE
ASESOR

Presentada a la Escuela de Postgrado de la Universidad Nacional Pedro Ruiz Gallo para optar el Grado académico de: MAESTRO EN INGENIERIA DE SISTEMAS CON MENCIÓN EN GERENCIA DE TECNOLOGIAS DE LA INFORMACIÓN Y GESTIÓN DEL SOFTWARE

APROBADO POR:

DR. LUIS ALBERTO DAVILA HURTADO
PRESIDENTE DEL JURADO

DR. REGIS JORGE ALBERTO DIAZ PLAZA
SECRETARIO DEL JURADO

M Sc. PILAR DEL ROSARIO RIOS CAMPOS
VOCAL DEL JURADO

Octubre, 2018

ACTA DE SUSTENTACIÓN DE TESIS 137

Siendo las 5:35 pm horas del día ONCE de ABRIL del año Dos Mil diecinueve, en la Sala de Sustentaciones de la Escuela de Postgrado de la Universidad Nacional Pedro Ruiz Gallo de Lambayeque, se reunieron los miembros del jurado, designados mediante Resolución Nº 0911 – 2018 – EPG de fecha 20 de abril del 2018., conformado por:

Dr. Luis Alberto Dávila Hurtado ..... PRESIDENTE (A)
Dr. Regis Jorge Alberto Díaz Plaza ..... SECRETARIO (A)
M. Sc Pilas del Rosario Rios Campos ..... VOCAL
..... ASESOR (A)

con la finalidad de evaluar la tesis titulada Optimización de la campaña de amnistía tributaria del centro de gestión tributaria de Chiclayo aplicando un modelo de predicción basado en el comportamiento de pago de los contribuyentes

presentado por el (la) tesista Carlos Alberto Inoñán González
sustentación que es autorizada mediante Resolución Nº 0400 – 2019 – EPG de fecha 27 de marzo de 2019

El Presidente del jurado autorizó el inicio del acto académico y después de la sustentación, los señores miembros del jurado formularon las observaciones y preguntas correspondientes, las mismas que fueron absueltas por el (la) sustentante, quien obtuvo 85 puntos que equivale al calificativo de Muy Bueno

En consecuencia el (la) sustentante queda apto (a) para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con mención en gerencia de tecnologías de información y gestión del software

Siendo las 6:25 horas del mismo día, se da por concluido el acto académico, firmando la presente acta.

PRESIDENTE

SECRETARIO

VOCAL

ASESOR

Observación: La asesora de tesis es la Dra. según resolución Nº 091 2018 – EPG la Dra. Emma Virginia Nobleclla Montenegro

R. Díaz P.

L. 07.06.2019

Dr. ARNULFO CIEZA RAMOS
Director Académico (e)

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO – ESCUELA DE POST GRADO – DIRECCIÓN ACADÉMICA – LAMBAYEQUE – PERÚ

NOTA: La existencia del acta en los libros de la Escuela de Posgrado de la Universidad Nacional Pedro Ruiz Gallo; ha sido verificada por la Sra. Moraima Vera Pozo, quien con su firma da fe de lo mencionado.

Sra. Moraima Vera Pozo
Trabajadora Administrativa

En el Acta de Sustentación se evidencia el proceso de sustentación de tesis. La misma que ha sido refrendada por el jurado conformado por presidente, secretario y vocal, más no, se registra la firma del asesor, cuya labor efectiva es durante el proceso de elaboración de tesis y su presencia en el acto de sustentación de la tesis es voluntaria. Por lo tanto, su ausencia no invalida el acto de sustentación.

El/la sustentante cumple con los requisitos para la emisión de su grado académico correspondiente.

Lambayeque, 07 de junio de 2019

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO ESCUELA DE POSGRADO DIRECCIÓN LAMBAYEQUE-PERU

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO ESCUELA DE POST GRADO DIRECCIÓN ACADÉMICA LAMBAYEQUE - PERU

Dra. OLINDA LUZMILA VIGO VARGAS
Directora

Dr. ARNULFO CIEZA RAMOS
Director Académico(e)

NOTA: La existencia del acta en los libros de la Escuela de Posgrado de la Universidad Nacional Pedro Ruiz Gallo; ha sido verificada por la Sra. Moraima Vera Pozo, quien con su firma da fe de lo mencionado.

Sra. Moraima Vera Pozo
Trabajadora Administrativa

DECLARACIÓN JURADA DE ORIGINALIDAD

Yo, Carlos Alberto Inoñán Gonzáles investigador principal, y Emma Virginia Noblecilla Montealegre, asesora del trabajo de investigación "OPTIMIZACIÓN DE LA CAMPAÑA DE AMNISTIA TRIBUTARIA DEL CENTRO DE GESTIÓN TRIBUTARIA DE CHICLAYO APLICANDO UN MODELO DE PREDICCIÓN BASADO EN EL COMPORTAMIENTO DE PAGO DE LOS CONTRIBUYENTES", declaramos bajo juramento que este trabajo no ha sido plagiado, ni contiene datos falsos. En caso se demostrará lo contrario, asumo responsablemente la anulación de este informe y por ende el proceso administrativo a que hubiere lugar. Que puede conducir a la anulación del título o grado emitido como consecuencia de este informe.

Lambayeque, 21 de junio del 2021

Nombre del investigador: Carlos Alberto Inoñán Gonzáles
Nombre de la asesora: Emma Virginia Noblecilla Montealegre

DEDICATORIA

La presente tesis se la dedico a toda mi familia, pilar fundamental en mi formación como profesional, por brindarme la confianza, consejos, oportunidad y recursos para lograrlo y a mis amigos por estar siempre en esos momentos difíciles brindándome su paciencia y compresión.

AGRADECIMIENTOS

La presente tesis es un reto académico y personal, y constituye un proceso de investigación en un campo en el que personas dentro de mi vida personal y profesional han intervenido con sus consejos, de una manera concreta y práctica, o demostrando su interés, cercanía y apoyo moral.

Por ello, quiero expresar mi profundo agradecimiento a:

Mis profesores de maestría, por su confianza en mi capacidad para llevar adelante esta investigación y haber compartido de diferentes maneras sus conocimientos.

A quienes han dirigido esta tesis, por su interés y orientación, sin ustedes esta investigación no hubieran sido posibles, por su dedicación y entrega inestimable.

A mis amigos Juan y Denis, sin cuyo apoyo y ayuda a la distancia no me hubiera sido posible superar los diversos obstáculos que se me presentaron a lo largo de todo este proceso.

TABLA DE CONTENIDOS

RESUMEN

La investigación tiene como objetivo determinar la probabilidad de incumplimiento de pago de los arbitrios municipales durante una campaña de amnistía tributaria, la campaña está dirigida a un conjunto de contribuyentes, quienes tienen deudas de arbitrios en cobranza morosa. El logro del objetivo fue alcanzado a partir de la aproximación a un modelo de regresión logística multivariada, aplicado a un conjunto de factores determinantes los cuales tienen como característica principal ser de naturaleza cuantitativa y cualitativa la cual influye en la probabilidad de incumplimiento de pago.

La metodología definida con sus ajustes necesarios y las variables seleccionadas para el proceso del diseño, finalmente muestran que como resultado de la investigación se pudo obtener un modelo de predicción para el incumplimiento de pago de los contribuyentes, a partir de doce variables de las que solo tres de ellas resultaron ser significativas con un nivel de confianza del 95%. No obstante debemos concluir que si bien con el modelo encontrado no se pudo lograr un buen ajuste estadístico; Sin embargo se pudo alcanzar una capacidad de pronóstico alta la cual supera el 92%; Por lo que se puede decir que estas variables no son suficientemente adecuadas o pertinentes para llevar a cabo un análisis de tales características, es así que debido a esta serie de dificultades presentadas, se recomienda determinar un conjunto de nuevas variables a estudiar con características que sean relevantes en el cálculo de la probabilidad del incumplimiento de pago de los arbitrios municipales de los contribuyentes de la Municipalidad Provincial de Chiclayo.

Palabras clave: Incumplimiento de pago, Contribuyente moroso, Probabilidad de incumplimiento de pago.

ABSTRACT

The objective of the investigation is to determine the probability of non-payment of municipal taxes during a tax amnesty campaign, the campaign is aimed at a group of taxpayers, who have debts of taxes in delinquent collection. The achievement of the objective was achieved from the approach to a multivariate logistic regression model, applied to a set of determining factors whose main characteristic is to be quantitative and qualitative in nature, which influences the probability of non-payment.

The methodology defined with its necessary adjustments and the variables selected for the design process, finally show that as a result of the investigation it was possible to obtain a prediction model for the non-compliance of taxpayers, from twelve variables of which only three of them were found to be significant with a confidence level of 95%. However, we must conclude that although the model found could not achieve a good statistical fit; However, it was possible to achieve a high prognostic capacity which exceeds 92%; Consequently it can be said that these variables are not adequate or relevant enough to carry out an analysis of such characteristics, it is thus that due to this series of difficulties presented, it is recommended to determine a set of new variables to study with characteristics that are relevant in the calculation of the probability of non-payment of the municipal taxes of the taxpayers of the Provincial Municipality of Chiclayo.

Key words: Breach of payment, Defaulter, Probability of default.

INTRODUCCIÓN

La optimización de la Campaña de Amnistía del Centro de Gestión Tributaria de Chiclayo mediante el uso de un Modelo que determine la probabilidad de incumplimiento de pago, es la respuesta a un problema que afronta la administración pública en el campo de los tributos recaudados por las municipalidades.

El problema del incumplimiento en el pago nos llevó a revisar la bibliografía existente y la recopilación de todas aquellas investigaciones que de alguna forma nos conduzcan a aclarar nuestra visión de la gestión del incumplimiento de pago y por lo tanto plantear nuevas y creativas formas de darle solución al mismo.

En el Centro de Gestión Tributaria no está definida aún una estrategia que permita conocer con claridad todos aquellos factores que conllevan al incremento del incumplimiento de pago del contribuyente.

Esta investigación consiste en determinar los factores asociados a los contribuyentes que influyen en el incumplimiento de pago de los arbitrios el cual se puede predecir mediante el uso de un modelo de Regresión Logístico Multivariado, el cual debe alcanzar una significancia estadística valida y permita definir a aquellos factores como atributos de riesgo en el no pago, considerando que el modelo logístico implementado, busca corroborar la influencia o no de dichos factores.

La justificación de la investigación se encuentra en que favorece la gestión del Centro de Gestión Tributaria de Chiclayo permitiéndole predecir con anticipación el resultado de las campañas de amnistía y de esta manera centrar su interés en la implementación de nuevas estrategias que consideren los factores resultantes del estudio.

Como parte de la investigación se establecieron los objetivos siguientes: como objetivo general obtener un Modelo de Predicción del Incumplimiento de Pago para la campaña de amnistía

tributaria del año 2018, como objetivos específicos: 1.- Determinar un grupo de factores que tienen influencia significativa sobre el incumplimiento de pago de arbitrios, 2.- Calcular la probabilidad de incumplimiento de pago de arbitrios con el algoritmo de regresión logística aplicado sobre un grupo de doce factores que son la entrada a el modelo, 3.- Determinar la validez y pertinencia del modelo resultante para pronosticar el incumplimiento de pago de los arbitrios municipales.

Se formuló la hipótesis general en la que los factores como: Número de predios, Autovaluo, Tipo de domicilio, Tipo de contribuyente, Calificación contributiva, Transferencia de venta de predios, Genero, Registro de Correo electrónico, Registro telefónico, Fiscalización de predio, Notificación de amnistía y Segmentación por tipo de pago no generan un modelo correctamente ajustado a la probabilidad de ser un contribuyente incumplido en el pago de arbitrios durante una Campaña de Amnistía Tributaria.

El estudio se estructuró de la siguiente forma:

En el capítulo I Contextualización, que permite ubicar el contexto del problema, mostrar el escenario actual en el que se desarrolla la investigación: describir donde se aplicará el modelo. Se presenta los antecedentes que son el resultado de investigaciones previas que permitieron determinar en qué momento se encuentran las investigaciones sobre incumplimiento de pago en general.

En el capítulo II Marco teórico se detalla el concepto de los términos más importantes para la investigación: incumplimiento de pago, regresión logística, segmentación de carteras.

En el capítulo III Metodología, que presenta la caracterización de la investigación y la validación del instrumento.

En el capítulo IV Resultados se presenta la evaluación de las hipótesis, la definición de las variables, el análisis estadístico y la contrastación de hipótesis.

La investigación presenta las conclusiones sobre el modelo resultante y recomendaciones para posteriores investigaciones.

CAP. I ANÁLISIS DEL OBJETO DE ESTUDIO

1.1. Contextualización

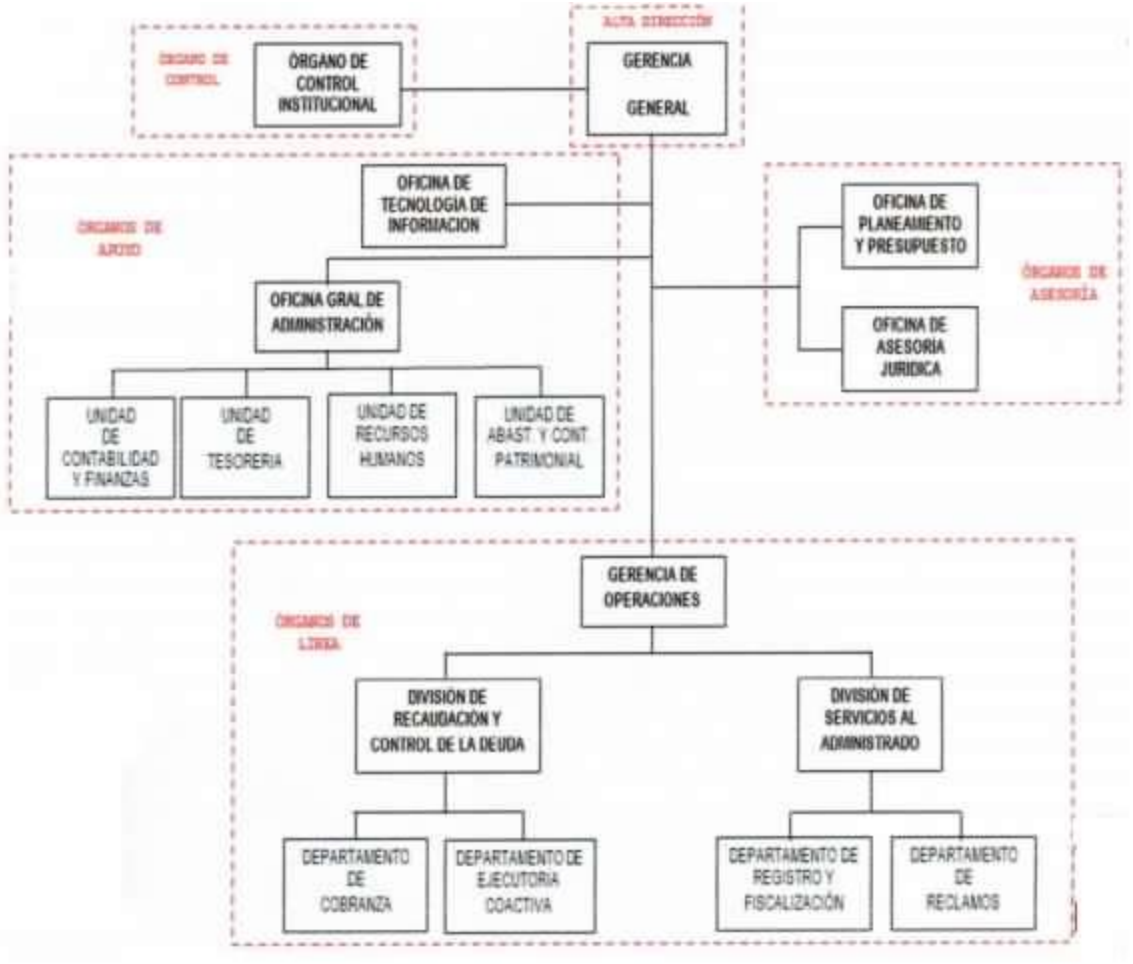
La alta dirección de la Municipalidad Provincial de Chiclayo llevo a cabo en mayo del 2003 una reorganización en el área de administración y recaudación tributaria local, razón por la cual se creó el Servicio de Administración Tributaria de Chiclayo, mediante Edicto Municipal N° 001-A-GPCH-2003, de fecha 13 de Mayo 2003, como un Organismo Público Descentralizado de la Municipalidad Provincial de Chiclayo, cuya labor principal es la administración, fiscalización y recaudación de los ingresos tributarios y no tributarios del Gobierno Provincial de Chiclayo, la cual no se ciñe a una mera labor de recaudación, sino que también lleva como cargo diversas actividades culturales y sociales.

Actualmente la denominación del Servicio de Administración Tributaria de Chiclayo se ha cambiado llamándosele en el momento de la investigación Centro de Gestión Tributaria de Chiclayo. Para nuestro propósito de estudio en adelante nos refiramos a él como CGT.

Proceso de Gestión de la Cobranza

La gestión de la cobranza en el CGT se lleva a cabo por la División de Recaudación y Control de Deuda a Cargo de la Gerencia de Operaciones.

Figura 1
Organigrama del Centro de Gestión Tributaria



FUENTE: Centro de Gestión Tributaria

La División de Recaudación y Control de la Deuda se encarga de planificar, organizar, dirigir, ejecutar y controlar las acciones de cobranza de la deuda tributaria y no tributaria cuyo proceso abarca desde la emisión de actos administrativos y otros documentos que el CGT necesite para la recuperación de la deuda. Dentro de las principales funciones que tiene se considera:

Planificar y ejecutar acciones de control de deuda y de fraccionamiento, de acuerdo con los programas, procedimientos y normatividad vigente.

Programar, supervisar y desarrollar las actividades que tienen que ver con la gestión y seguimiento de la cobranza pre-coactiva y coactiva de las deudas de tipo tributario y no

tributario.

Evaluar el nivel en que son efectivas las acciones de gestión de cobranza y determinar las acciones necesarias para subir el nivel de recaudación.

Procedimiento de Administración y Gestión de Carteras

Consta de cuatro etapas:

1. Establecer las herramientas y recursos necesarios para la gestión

   En esta etapa el jefe del Departamento de Cobranza organiza dos equipos que gestionaran las carteras de PRICOS y MEPECOS. Para lo cual contaran con los recursos y herramientas necesarias.

   Herramientas

   - Control de productividad de los gestores
   - Control del registro de resultados
   - Control del cumplimiento de visitas
   - Control del Fraccionamiento

   Recursos

   - Abastecimiento de material
   - Gestión telefónica
   - Orientación al contribuyente
   - Impresión de estados de cuenta

2. Depuración de la base de contribuyentes

   Esta etapa inicia solicitando a la Oficina de Tecnologías de Información la base de contribuyentes morosos actualizada, dicha base será sincerada en esta etapa filtrando de la

misma los contribuyentes que no cumplen los requisitos necesarios para el cobro de la deuda, entre esos tenemos:

- Direcciones inubicables
- Contribuyentes con reclamos en curso
- Contribuyentes con predios en proceso de fiscalización

Como resultado de esta fase se tiene una base de contribuyente de cobranza dudosa que no ingresaran a la gestión por no tener sincerada su deuda tributaria la cual es demasiado difícil de recuperar.

3. Segmentación de carteras

   La base diferencial resultante de la etapa anterior será segmentada en tres grupos de contribuyentes los cuales serán etiquetados como Principales, Medianos y Pequeños en función del monto de base imponible afecta para el año fiscal en curso.

   La segmentación de cartera resultante debe ser aprobada por la Gerencia de Operaciones después de lo cual dicho atributo será actualizado por la Oficina de Tecnologías de información en las bases de datos de contribuyentes actual.

   Finalmente, en esta etapa la cartera de Medianos y Pequeños contribuyentes se reagrupa para su gestión conjunta en la cartera de MEPECOS las cuales serán zonificadas por ubicación geográfica para su gestión.

4. Gestión de cobranzas de PRICOS y MEPECOS

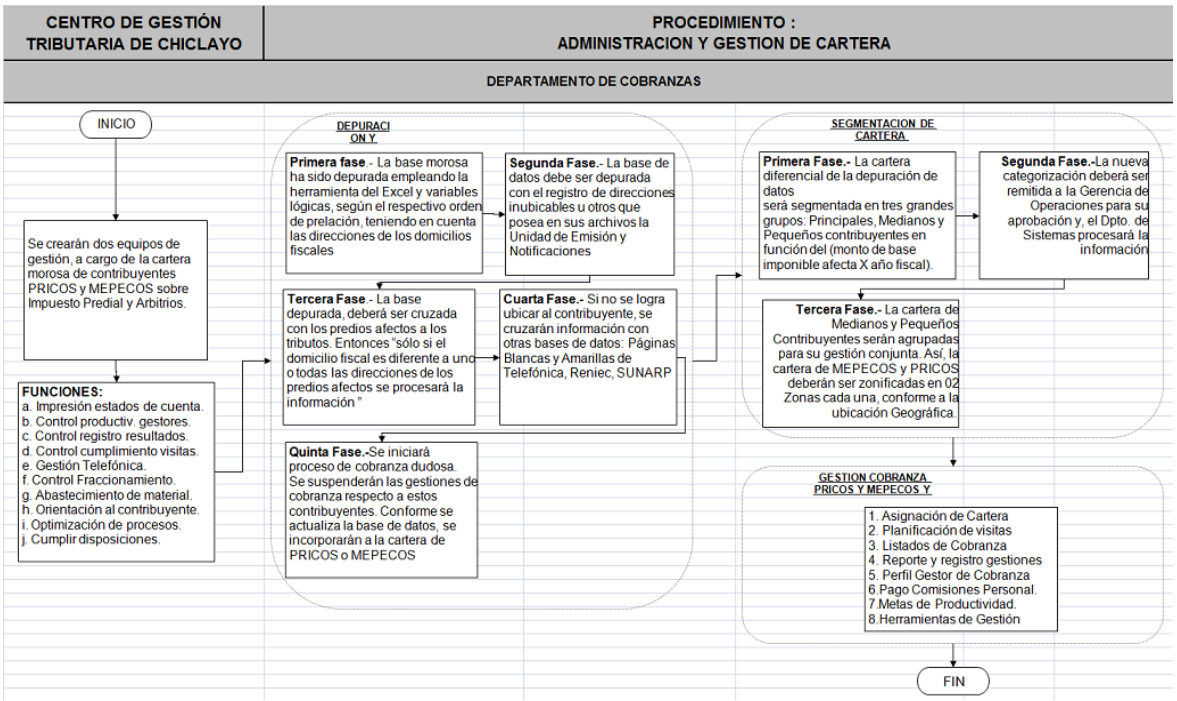
   Esta etapa tiene como principal característica ser operativa y abarca un conjunto de actividades permiten la cobranza in situ de la deuda de cada contribuyente etiquetado en las carteras.

Dentro de estas actividades tenemos:

- Asignación de carteras
- Planificación de vistas a los contribuyentes
- Listados de cobranzas para los gestores
- Reportes y registro de gestiones
- Establecimiento de perfiles de los gestores de cobranza
- Pagos de comisiones al personal
- Establecimiento de las metas de productividad

En la figura 2 se muestra un resumen del proceso de Administración y Gestión de Cartera.

Figura 2
Procedimiento: Administración y Gestión de Cartera



FUENTE: Elaborada por el Centro de Gestión Tributaria

Del proceso de Administración y Gestión de carteras hemos determinado que no ha variado

significativamente desde su implementación original el año 2004.

Identificándose como problema principal que la etapa de segmentación de carteras clasifica a los contribuyentes utilizando solo un atributo “el monto de la base imponible por año fiscal” no considerando un conjunto de atributos que caen en las siguientes categorías:

- Geográficos
- Demográficos
- Psicográficos
- Comportamiento

Considerar el listado de atributos previo permitirá que la etapa de segmentación de carteras incorpore nuevos atributos que faciliten establecer un perfil de contribuyente sobre el cual las estrategias de cobranzas se ajustaran para extraer el máximo rendimiento posible de las carteras de cobranzas que se gestione. Para esta nueva etapa de segmentación es necesario implementar un proceso que posibilite extraer patrones de los datos almacenados del contribuyente y nos permita establecer perfiles más específicos.

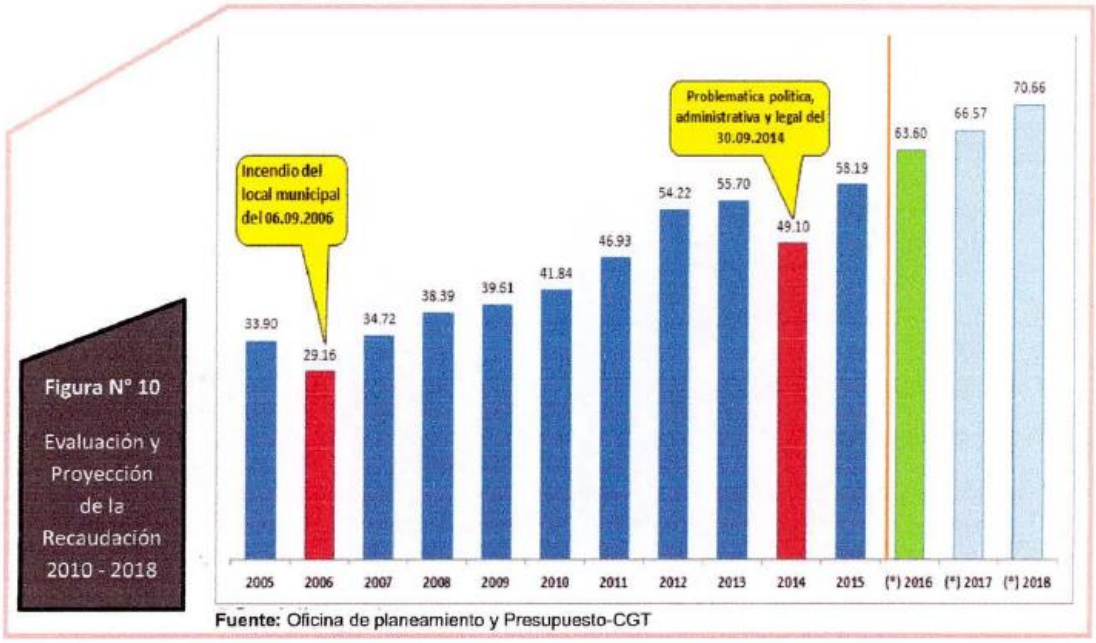
Análisis de la Evolución de la Recaudación

En el contexto institucional de: "Ser líder y gestionar con eficiencia la recaudación de los ingresos tributarios y no tributarios de la Municipalidad Provincial de Chiclayo"; la alta dirección de la Central de Gestión Tributaria (CGT) está comprometida con seguir incrementando la recaudación de las obligaciones de los contribuyentes del Distrito de Chiclayo proyectándose para el año 2018 superar los 70 millones de nuevos soles. Permitiendo así alcanzar una estabilidad económica-financiera al CGT y a la actual administración municipal.

En el siguiente grafico se puede apreciar que desde el año 2005 al 2015 la recaudación tiene una tendencia positiva. De igual manera para el periodo 2016 - 2018 las proyecciones de la

recaudación se esperan sigan creciendo de manera positiva.

Gráfico 1
Evaluación y Proyección de la Recaudación 2010 – 2018



FUENTE: Elaboración. Oficina de planeamiento y presupuesto CGT

Como se puede observar en el cuadro anterior, que tiene el monto proyectado de la recaudación de los años 2016, 2017 y 2018, de la información recopilada del Plan Operativo Institucional del año 2017; En comparación con lo que se observa en el siguiente cuadro, elaborado al cierre del periodo fiscal del año 2017, extraído del documento *Índices de recaudación al mes de Diciembre del 2017*.

Gráfico 2

Evaluación de la recaudación acumulada al mes de diciembre años 2003 - 2017

FUENTE: Elaboración. Oficina de planeamiento y presupuesto CGT

Se ha determinado que existe un déficit entre lo recaudado y lo proyectado de 7.38% y 13.09% para los años 2016 y 2017 respectivamente. Dicho déficit podemos observar se ha venido incrementando.

1.2. Surgimiento del problema

Modelo Logit para determinar los factores socio-económicos que influyen en el incumplimiento de pago del impuesto predial del Distrito de Piura. Cuyo propósito es encontrar un modelo Logit Multivariado que nos permita identificar los factores socioeconómicos de incumplimiento en base a información recolectada de los contribuyentes del distrito de Piura. (Chunga Chully & Reyes Pintado, 2015)

El Modelo predictivo de fuga de clientes incorporando minería de datos para una compañía de telecomunicaciones en Chile: En este trabajo se implementa un modelo de fuga de clientes para

una empresa de telecomunicaciones que compite en dos mercados, Concepción y Temuco, de Chile. Se emplean como metodologías el análisis de clúster para generar perfiles de clientes fugados y la técnica de regresión logística multivariable para obtener un modelo de ocurrencia de fuga de servicios. La base de datos incluyó productos contratados, variables socio demográfico, sistemas de pago, número y tipo de reclamos, entre otros. Se generan modelos de regresión logística multivariable para cada una de las localidades. (Jélvez Camaño, Moreno Echevarria, Ovalle Retamal, Torres Navarro, & Troncoso Espinoza, 2014)

Factores determinantes para la probabilidad de incumplimiento para un microcrédito en una entidad microfinanciera en Perú, un acercamiento bajo el modelo de regresión logística binaria. Esta investigación busca medir la probabilidad de incumplimiento de un microcrédito mediante una función logística binaria, que estudia la relación entre factores determinantes cuya naturaleza es cuantitativa y cualitativa para determinar la probabilidad de incumplimiento para un microcrédito (Calixto Salazar & Casaverde Carranza, 2011)

Clasificación de grandes conjuntos de datos vía máquinas de vectores de soporte y aplicaciones en sistemas biológicos: Analiza varios modelos que toman como base máquinas de vectores de soporte con el fin de mejorar el desempeño en múltiples aplicaciones del mundo real; debido al elevado coste computacional de este modelo en conjunto de datos grandes. (Cervantes Canales, 2009)

Predicción de fuga de clientes para una institución financiera mediante Support Vector Machine: Se presenta un modelo predictivo implementado para etiquetar a los clientes con tendencias a la fuga en un banco con el objetivo de hacer más efectivas las políticas comerciales

de retención, hacer más eficiente la asignación de recursos y mejorar la relación con los clientes al identificar los principales puntos de deficiencia del servicio (Miranda, Rey, & Weber, 2005)

Reconocimiento de patrones de evasión en el sistema de administración tributaria empleando tecnología Datamining: Se muestra un modelo predictivo implementado para identificar a los contribuyentes con señales de evasión. Aplicando redes neuronales. (Vilcapoma, 2003)

CAP. II MARCO TEÓRICO

2.1. Teorías

2.1.1. Arbitrios

Es el tributo obligatorio que genera la prestación efectiva y/o potencial por el Estado de un servicio público individualizado en el contribuyente. Es decir, su obligación de pago da soporte y financia los servicios públicos que presta la Municipalidad. (Congreso, 2004).

Los sujetos que son pasivos a el pago de esta tasa son aquellos propietarios de inmuebles urbanos ubicados en la jurisdicción del distrito de Chiclayo.

2.1.2. Evasión Tributaria

(Sampaio Dória, 1971), la definió como cualquier acto que tiende a evitar, disminuir o retrasar el pago de una obligación de índole tributario.

Además, el las categoriza de la siguiente manera:

- Abstención de incidencia: cuando el contribuyente evita llevar a cabo actos que generen obligaciones tributarias y de esta manera evita incurrir en falta con la ley.
- Transferencias económicas: se da por una desviación económica del tributo de derecho hacia el de hecho, la cual tampoco sería algo que incurra en la ilegalidad.
- Evasión por inacción: está supeditada a si es intencional, por una acción no consciente y de manera involuntaria, o no intencional, debido al escaso conocimiento del contribuyente y a la elevada complejidad del tributo generado.
- Evasión ilícita: Es el tipo de evasión más conocida, iniciada de manera consiente y voluntaria del deudor, quien utiliza cualquier medio ilícito para evadir el cumplimiento tributario. Constituyendo en si este tipo de operaciones en fraudulentas o de simulación.
- Evasión Lícita: Esta es el acto que conlleva el alejar, disminuir o prolongar la ejecución del hecho que origino el tributo realizado por una serie de procesos lícitos.

2.1.3. Definición de incumplimiento

Según (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010), la definición del incumplimiento de pago no debe hacerse de forma apresurada ya que es necesario identificar todo retraso que afecte a la organización, el cual debe cumplir las siguientes condiciones:

- El atraso debe ser identificado de manera real y no simplemente estimado siguiendo un cronograma de fechas de pago.
- El atraso ha de presentarse en, al menos, una cuota de amortización.
- El atraso deber tener su equivalente en costes monetarios en los que se incide cuando se ejecuta el seguimiento y gestión de la deuda atrasada.

2.1.4. Modelos de elección discreta

Los modelos de elección discreta son caracterizados por que la variable dependiente tiene carácter cualitativo (no métrica). Además de lo anterior están relacionados con el análisis discriminante. En el momento actual existe una mayor incidencia en el uso de este tipo de modelos en razón a que se necesita definir menos supuestos, circunstancia que permite, en mayor medida obtener mejores resultados.

En estadística, la regresión logística es una categoría de análisis de regresión que se utiliza para pronosticar el resultado de una variable categórica (una variable que puede admitir un número reservado de categorías) en función de las variables independientes o predictoras. Es útil para configurar la probabilidad de un evento ocurrido en función de otros factores. El análisis de regresión logística se encuadra en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que hace uso como función de enlace la función logit. Las probabilidades que delinean el posible resultado de un único examen se modelan, como una función de variables explicativas, para lo cual utilizan una función logística.

La regresión logística es usada ampliamente por las ciencias médicas y sociales. Otros nombres

para regresión logística usados en varias áreas de aplicación incluyen modelo logístico, modelo logit, y clasificador de máxima entropía. La Regresión Logística es una técnica estadística multivariante que nos permite aproximar la relación que existe entre una variable dependiente no métrica, en particular dicotómica y un grupo de variables independientes métricas o no métricas.

2.1.4.1. Regresión Logística Multinomial

Es un modelo de regresión logística cuya variable dependiente tiene más de dos categorías. La respuesta puede ser nominal u ordinal. A la vez, las variables explicativas pueden tener naturaleza categórica o cuantitativa. En los modelos de regresión multinomial se ha de suponer que los recuentos de las categorías de Y tienen una distribución multinomial. Esta distribución será una generalización de la distribución binomial. (Marín Diazaraque, 2011)

Modelos Logit para respuestas nominales

(Marín Diazaraque, 2011) El número de categorías de Y se denota como J donde $\{\pi_1, \ldots, \pi_J\}$ son las probabilidades de las distintas categorías tal que $\sum_j \pi_j = 1$

Se parte de n observaciones independientes que se distribuyen en las J categorías. La distribución de probabilidad del número de observaciones de las J categorías sigue una distribución multinomial. Esta distribución modela la probabilidad de cada una de las posibles maneras en que n observaciones pueden repartirse entre las J categorías. Al ser la escala de medida nominal, el orden entre las categorías no tiene relevancia. Se toma una categoría como respuesta base, por ejemplo, la última categoría (J), y se define un modelo logit con respecto a ella:

$$\log(\pi_j/\pi_J) = \alpha_J + \beta_J x$$

Donde j = 1…, J − 1.

El modelo tiene J − 1 ecuaciones con sus propios parámetros, y los efectos varían con respecto a la categoría que se ha tomado como base.

Cuando J = 2, el modelo equivale a una única ecuación log (π_1/π_2) = logit(π_1) y se obtiene el modelo de regresión logística estándar.

La ecuación general logit con respecto a la categoría base J determina también los logits para cualquier pareja de categorías. Así, si consideramos dos categorías cualesquiera a y b,

$$\log(\pi_a/\pi_b) = \log\left(\frac{\frac{\pi_a}{\pi_j}}{\frac{\pi_b}{\pi_j}}\right) = \log(\pi_a/\pi_J) - \log(\pi_b/\pi_J) = (\alpha_a + \beta_a x) - (\alpha_b - \beta_b x) =$$

$$(\alpha_a - \alpha_b) + (\beta_a - \beta_b)x$$

De este modo, la ecuación para las categorías a y b tiene también la forma $\alpha + \beta_x$ donde

$\alpha = (\alpha_a - \alpha_{b)}$ y $\beta = (\beta_a - \beta_{b)}$

Entre los modelos de elección discreta más conocidos se tienen:

Cuadro 1
Clasificación de los modelos de elección discreta

Modelo	**Descripción**
Modelos dicotómicos o binomiales	Cuando la variable dependiente puede adoptar solo dos modalidades o categorías. Las dos modalidades deben ser mutuamente excluyentes.
Modelos Multinomiales	Cuando la variable dependiente puede adoptar más de dos modalidades que son diferentes, exhaustivas y mutuamente excluyentes.
Modelos Ordenandos	Cuando la variable dependiente puede adoptar más de dos modalidades, que son también, diferentes, exhaustivas y mutuamente excluyentes pero que, a diferencia de los

	multinomiales puede determinarse un orden 'es decir' es una variable ordinal.
Modelos Loglineales	Estos modelos se utilizan para examinar tablas de contingencia de dos o más dimensiones.

2.1.5. Modelos de probabilidad no lineal

La estimación e interpretación de los modelos probabilísticos lineales conlleva un conjunto de problemas que han conducido a la búsqueda de diferentes modelos alternativos que van a permitir aproximaciones más fiables de las variables dicotómicas. Para impedir que la variable endógena estimada pueda ubicarse fuera del rango [0; 1], las alternativas de que se dispone son utilizar modelos de probabilidad no lineales, donde la función de especificación a utilizar asegure un resultado en la estimación comprendido en el rango [0; 1]. Las funciones de distribución garantizan que se cumpla este requisito, ya que son funciones continuas que consideran valores comprendidos entre 0 y l.

2.1.5.1. Especificación de los modelos de elección discreta Logit y Probit

Teniendo en cuanta que una función de distribución nos asegura que el resultado de la estimación esté limitado entre 0 y 1, como principio las alternativas pueden ser varias, siendo las más frecuentes la función de distribución logística, que tiene como resultado al modelo Logit, y la función de distribución de la normal tipificada, que tiene al modelo Probit como resultado. Como los modelos Logit y los Probit se relacionan, por tanto, la variable endógena Y_i con las variables explicativas X_i se relacionan a través de una función de distribución.

Dado a que el modelo Logit, utiliza la función logística, por lo que la especificación de esta categoría de modelo queda como sigue:

$$Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}}$$

Para el modelo Probit en cambio utilizara como función de distribución la normal tipificada, con lo que dicho modelo queda especificado a través de la siguiente expresión

$$\gamma_i = \int_{-\alpha}^{\alpha + \beta X_I} \frac{1}{(2\pi)^{1/2}} e^{\frac{x^2}{2}} dz + \epsilon_i$$

Donde la variable z es una variable "muda" de integración con media cero y varianza uno.

Dada la similitud que se encuentra entre las curvas de la normal tipificada y de la logística, los resultados obtenidos por ambos modelos no guardan gran diferencia, siendo las mismas operativas, y que se presentan por la complejidad en el cálculo de la función de distribución normal frente a la logística, ya que la primera solo puede calcularse en forma de integral. La reducida complejidad de manejo que caracteriza al modelo Logit es lo que ha beneficiado su amplio uso en la mayoría de los estudios empíricos.

De la misma forma que el Modelo Lineal de Probabilidad, el Modelo Logit se puede interpretar en términos probabilísticos, es decir, puede servir para medir la probabilidad de que ocurra el evento objeto de estudio (Y;= 1). En cuanto a la interpretación de los parámetros determinados en un modelo Logit, el signo de los parámetros indica la dirección en que se mueve la probabilidad cuando aumenta la variable explicativa correspondiente, sin embargo, el valor del parámetro no coincide con la medida de la variación en la probabilidad (como si ocurría en el MLP). En el caso de los modelos Logit, al suponer una relación no lineal entre las variables explicativas y la probabilidad de ocurrencia del evento, cuando se incrementa en una unidad la variable explicativa los incrementos en la probabilidad no son siempre iguales ya que estos dependerán del nivel original de la misma.

2.1.6. La Ecuación Logística

La regresión logística es una herramienta estadística de análisis bivariado o multivariado, cuyo uso puede ser tanto explicativo como predictivo. Dicha herramienta se utiliza cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un grupo de m variables predictoras o independientes, cuya naturaleza puede ser cuantitativas (que se denominan covariables o covariadas) o categóricas. Cuando son categóricas, se requiere que sean convertidas en variables ficticias o simuladas ("dummy"). El propósito del análisis es:

- Predecir la probabilidad de que a alguien le ocurra cierto evento: por ejemplo, "estar enfermo" = 1 o "no estarlo" = 0; "ser rico" = 1 o "no ser rico"= 0; "graduarse como médico" =1 o "no graduarse"= 0.
- Determinar qué variables tienen mayor peso para incrementar o disminuir la probabilidad de que a alguien le ocurra el evento en cuestión.

Esta determinación de probabilidad de ocurrencia del evento a cierto sujeto, así como la determinación de la influencia que cada una de las variables independientes tienen en esta probabilidad, se basan en las características que muestran los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

La regresión logística binaria (a la que desde ahora solo se llamara regresión logística) se aplica con una variable dependiente dicotómica, donde la variable dependiente no contiene valores de datos sin procesar, pero en cambio representa la oportunidad de que en un evento de interés se presente. En términos generales la ecuación de la regresión logística es:

$$\ln\left(\frac{P_i}{1-P_I}\right) = \alpha + \beta_K X_{ki} + \varepsilon_i$$

Donde en la parte derecha se encuentran los términos estándar para las variables independientes y la intercepción es una ecuación de regresión, sin embargo, en la parte izquierda está el logaritmo natural de la oportunidad y la cantidad ln(Odds) llamada logit. En principio este valor puede variar de menos a más infinito, por lo que se elimina el problema de la predicción fuera de los límites de la variable dependiente. La oportunidad está vinculada con la probabilidad por:

$$\boldsymbol{Odds = \frac{P_i}{1 - P_i}}$$

Se puede ver que hay una relación lineal con las variables independientes en la regresión logística, pero es lineal dentro del logaritmo natural de la oportunidad y no en las probabilidades originales dado que nuestro interés está enfocado en la probabilidad de un evento, por ejemplo, el código más alto de una variable dicotómica o categoría de interés, la ecuación logística se puede transformar en otra ecuación en la probabilidad, entonces tiene esta forma:

$$Prob\ (evento\ Y_i = 1) = \frac{1}{1 + e^{-(\alpha + \beta_{ki} X_{ki})}}$$

Dónde: Prob ($y = \frac{1}{x}$) es la probabilidad de qué y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X:

X_{ki}: es un conjunto de k covariables que forman parte del modelo.

α: es la constante del modelo o término independiente.

β_{ki}: Los coeficientes de las covariables.

Esta ecuación no se puede determinar con el método de mininos cuadrados, en su lugar, los parámetros del modelo se calculan con la técnica de máxima verosimilitud. Derivamos los coeficientes que hacen que nuestros valores observados sean los más probables para el grupo de las variables independientes, esto se hace a través de iteraciones internas realizadas por los programas estadísticos. La probabilidad resultante es el punto de partida para la clasificación de cualquier elemento de la muestra en uno de los grupos definidos dentro de la variable de respuesta.

Los modelos de regresión logística binaria atraen el mayor interés ya que la gran parte de las circunstancias estudiadas en el campo de la investigación responden a este modelo (presencia o no de enfermedad, éxito o fracaso, etc.). Como se ha observado, la variable dependiente será una variable dicotómica que se codificará como 0 o l (respectivamente, "ausencia" y "presencia"). Este punto de la codificación de las variables no es banal (influye en la forma en que se realizan los cálculos matemáticos), y habrá que tenerlo muy en cuenta cuando se emplean programas estadísticos que no recodifican automáticamente las variables cuando éstas se encuentran codificadas de una manera diferente (por ejemplo, el uso frecuente de 1 para la presencia y -1 ó 2 para la ausencia).

La ecuación de partida en los modelos de regresión logística se denomina distribución logística (Silva Ayçague, 1995)

2.1.7. Elementos del Análisis de Regresión Logística

Cuando se hace un análisis de regresión logística se tiene dos objetivos generales:

1. Determinar el efecto de un grupo de variables en la probabilidad, además del efecto de las variables tomadas individualmente sobre el resultado general del modelo. Esto significa la

ejecución de medir la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva tácito también aclarar la existencia de interacción y confusión entre covariables con respecto a la variable dependiente (es decir, conocer los Odds ratio para cada covariable).

2. Alcanzar la más alta precisión predictiva posible con un determinado grupo de variables predictoras seleccionadas a partir de la significancia obtenida por cada una de ellas. También se supone como consecuencia la acción de clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la aparición de determinadas covariables.

Estos dos objetivos no son incompatibles, pero si tienen un enfoque más específico. Quienes están interesados en la teoría y en los efectos causales se preocupan más en el primer objetivo, mientras lo que están interesados en la predicción y la aplicación funcional del modelo sobre una base de datos en el contexto real se fijaran más en el segundo objetivo. De lo que no cabe duda es que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación, de ahí su amplia utilización. El objetivo esencial que se alcanza con esta técnica es el de modelar cómo influye en la probabilidad de aparición de un evento, habitualmente dicotómico, la presencia o ausencia de diversos factores y el valor o nivel de los mismos.

También puede ser usada para calcular la probabilidad de aparición de cada una de las posibilidades de un evento con más de dos categorías. (Silva Ayçague, 1995)

2.1.8. Supuestos de la Regresión Logística

(Silva Ayçague, 1995) La regresión logística necesita que:

1. Las variables independientes sean de intervalos, de razón o dicotómicas.
2. Que se incluya a todas las predictoras relevantes, que no incluya variables irrelevantes y que tengan una relación lineal entre sí.

3. El valor esperado del error es 0.

4. Que no haya autocorrelación.

5. Que no haya correlación entre el error y las variables independientes.

6. Que no haya multicolinealidad entre las variables independientes.

2.1.9. Contraste y Validación de Hipótesis

Cuando se trabaje con observaciones repetidas la contrastación y validación del modelo estimado sigue la misma metodología que la utilizada en el análisis de regresión tradicional, por lo que referimos a éste para profundizar en este asunto. Dado que nos encontramos en el caso de no disponer de observaciones repetidas, la etapa de contrastación y validación del modelo estimado por máxima-verosimilitud se lleva a cabo aplicando los estadísticos específicos que se desarrollan a continuación.

2.1.9.1. Significatividad estadística de los parámetros estimados

(Silva Ayçague, 1995) La distribución del estimador del parámetro β es aproximadamente $N(\beta; \sigma_\beta)$. En tal circunstancia, se puede construir un intervalo de confianza del parámetro estimado, para probar si dicho valor es significativamente distinto de cero de forma individual. El contraste a efectuar quedaría definido como:

$\mathbf{H_0}$: $\boldsymbol{\beta = 0}$ El parámetro es igual a cero

H1: $\boldsymbol{\beta \neq 0}$ El parámetro es distinto a cero

El intervalo de confianza provee un rango de posibles valores para el parámetro, por lo que, si el valor estimado no corresponde a dicho intervalo, se deberá rechazar la hipótesis nula. El intervalo quedaría definido como:

$$\beta - Z_{\frac{\alpha}{2}}\, \sigma_\beta \;\leq\; \beta \;\leq\; \beta + Z_{\frac{\alpha}{2}}\, \delta_\beta$$

Donde α es la probabilidad de que el verdadero valor del parámetro β se encuentre fuera del intervalo, y z es el valor tabular de la distribución N(0; l) que deja a su derecha una probabilidad igual a $\alpha/2$.

A partir de la expresión anterior se puede determinar un rechazo de la hipótesis nula cuando:

$$\left|\frac{\beta}{\delta_\beta}\right| \geq Z_{\alpha/2}$$

2.1.9.2. Medidas de Bondad de Ajuste del Modelo

Se considera que cualquier variable dependiente de otra u otras variables, toma valores de acuerdo a las variables de las que depende. De otro lado, dicha variable dependiente irá adoptando valores siguiendo o describiendo una determinada distribución de frecuencias; es decir, adopta los valores que tienen las variables independientes, si el experimento se repite muchas veces, la variable dependiente tomará para un grupo de variables independientes un determinado valor, y la probabilidad de ocurrencia de aquel valor vendrá dado por una distribución de frecuencias concreta: una distribución normal, una distribución binomial, una

distribución hipergeométrica, etc. (Silva Ayçague, 1995)

Estadístico de Wald:

Nos permite determinar la significancia de los coeficientes, está definido como el vector matriz de los coeficientes estimados del siguiente modo según las Hipótesis y donde se busca validar la proposición de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza l. Su valor para un coeficiente determinado este dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación nos señala que dicho coeficiente es diferente de 0 y vale la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald nos puede brindar falsas ausencias de significación (es decir, se incrementa el error tipo II). Tampoco es recomendable su uso si se están empleando variables de diseño. (Silva Ayçague, 1995)

$$H_0: \beta_i = 0, \forall i$$

$$H_1: Al\ menos\ un\ \beta_i \neq 0$$

Índice de Cociente de Verosimilitudes:

La función de verosimilitud sirve de base para obtener un estadístico, que tiene cierta similitud con el coeficiente de determinación determinado en la estimación lineal, conocido "índice de cociente de verosimilitudes". Este estadístico contrasta el valor de la función de verosimilitud de dos modelos; el primero que corresponde al modelo estimado que incluye todas las variables independientes (modelo completo) y el segundo que sería el del modelo cuya única variable independiente es la constante (modelo restringido). El estadístico, también conocido como R^2 de McFadden ya que fue propuesto por McFadden en 1974, se define como:

$$RV = ICV = 1 - \frac{\log L}{Log\ L(0)}$$

Donde L representa el valor de la función de verosimilitud del modelo completo (el estimado con todas las variables independientes) y L(0) corresponde al valor del modelo restringido (el que incluye únicamente en la estimación el término constante).

Se trata de ir comparando cada modelo que se consigue de eliminar de forma aislada cada una de las covariables en contraste con el modelo completo.

En este caso cada estadístico R.V. sigue una X^2 con un grado de libertad (no se supone normalidad). La ausencia de significación conlleva que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no contribuye nada al mismo. Esta prueba no supone ninguna distribución concreta, por lo que es la más aconsejada para estudiar la significación de los coeficientes.

El ratio calculado tendrá valores comprendidos entre 0 y 1 de manera que los Valores próximos a 0 corresponderán a cuando L(0) sea muy parecido a L, escenario en el que nos encontraremos cuando las variables incluidas en el modelo sean poco significativas, es decir, el cálculo de los parámetros P no mejora el error que se comete si dichos parámetros se igualaran a 0. Por lo que en este caso la capacidad explicativa del modelo será muy reducida.

Cuanto mayor capacidad explicativa tenga el modelo, mayor será el valor de L sobre el valor de L(0), y más se acercara el ratio de verosimilitud calculado al valor l. (Silva Ayçague, 1995)

El estadístico X^2 de Pearson:

Para determinar la bondad del ajuste igualmente se utilizan medidas del error que cuantifican la diferencia entre el valor observado y el estimado.

En concreto, para contrastar la hipótesis nula de que:

$$H_0: Y_i = \widehat{P}_i \, ; lo\ que\ equivale\ a\ H_0: Y_i - \widehat{P}_i = e_i = 0$$

Se construye un estadístico que agrupa los residuos estandarizados o de Pearson del modelo Logit, que se determinan como la diferencia entre el valor observado de la variable respuesta y el estimado, dividido por la estimación de la desviación típica, ya que la esperanza es nula.

Este estadístico se asemeja a la suma de cuadrados de los residuos del modelo de regresión convencional. El ajuste del modelo mejorará cuanto más cerca esté el valor del estadístico de cero. Para conocer a partir de qué valor puede considerarse el ajuste como aceptable es necesario saber la distribución del estadístico. Éste estadístico, bajo la hipótesis nula, se distribuye como una chi-cuadrado con (n - k) grados de libertad, por lo que su valor se compara con el valor teórico de las tablas de la chi-cuadrado para contrastar la hipótesis nula. Si el valor calculado es mayor al valor teórico se rechazará la hipótesis nula lo que es equivalente a decir que el error incurrido es significativamente distinto de cero, es decir, se trata de un mal ajuste. (Silva Ayçague, 1995)

Porcentaje de aciertos estimados del modelo:

Otra de las vías utilizadas para determinar la bondad de un modelo Logit es predecir con el modelo los valores de la variable dependiente Yj de tal manera que Yj = 1 si $\beta_i > c$ ó Yj = O si $\beta_i > c$. Generalmente, el valor que se asigna a c para calcular si el valor de la predicción es igual a 1 o a 0 es de 0.5, puesto que parece lógico que la predicción sea 1 cuando el modelo

dice que es más probable obtener un 1 que un 0.

No obstante, la elección de un umbral igual a 0.5 no siempre es la mejor alternativa. En el caso en que la muestra se encuentren desequilibrios entre la cantidad de unos y el de ceros la elección de un umbral igual a 0.5 podría llevar a no predecir ningún uno o ningún cero.

Modificando el valor del umbral disminuirá siempre la probabilidad de un error de un tipo y se incrementará la probabilidad del otro tipo de error. Por lo que el valor que debe tomar el umbral depende de la distribución de datos en la muestra y de la importancia relativa de cada tipo de error.

Una vez se ha escogido el nivel del umbral, y dado que los valores reales de Y¡ son conocidos, basta con enumerar el porcentaje de aciertos para decir si la bondad del ajuste es elevada o no. A partir de esta enumeración se puede construir el siguiente cuadro de clasificación:

Cuadro 2
Cuadro de clasificación de aciertos

		Valor Real de Y_i	
		$Y_i = 0$	$Y_i = 1$
Predicción de $\widehat{P}_i$	$\hat{P}_i$ < c	P_{11}	P_{12}
	$\hat{P}_i$ > c	P_{21}	P_{22}

Donde P_{11} y P_{22} corresponderán a predicciones correctas (valores 0 bien predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P_{12} y P_{21} corresponderán a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal predichos en el segundo caso). (Silva Ayçague, 1995)

Precisión en la Predicción:

Un indicador de lo bien que se desempeña el modelo está en su capacidad de clasificar a los casos con precisión en las dos categorías definidas por la variable predicha, la precisión predictiva global y las precisiones especificas se obtienen en el cuadro de porcentajes a partir de una división del número de casos estimados de forma correcta sobre el total de casos clasificados para ambas categorías que se conocen antes de iniciar el trabajo de regresión logit. Estos valores son importantes para producir respuestas inmediatas en cuanto a la buena predicción que se consigue con el modelo, y debe ser comparada inmediatamente con las pruebas de significancia para verificar si ambos objetivos son compatibles o las diferencias son notorias, bajo el concepto inicial de la falta de correspondencia entre el ajuste del modelo y los estadísticos de verosimilitud, o la significancia de las variables individuales y la habilidad predictiva del mismo. Dado que hallar un modelo significativo no es razón suficiente de tener una elevada predictibilidad. A los valores de porcentajes particulares para la tabla de aciertos se les conoce como especificidad que se hace en función de la categoría más común o específica y para la categoría menos común se le conoce como sensibilidad, por ser la categoría sensible o de interés sobre el estudio y utilización del modelo logit. (Silva Ayçague, 1995)

Realizando predicciones

Con los coeficientes de regresión se puede hacer predicciones sobre los valores de casos individuales, entonces vamos a calcular la probabilidad de la categoría de interés en el estudio, sustituyendo los valores apropiados individuales en cada variable para dicho caso individual en la función matemática del modelo ya conocido, este resultado se puede reflejar en términos de oportunidad con la consideración exponencial que de elevar dicha probabilidad al valor (e) como se discutió para los valores de los coeficientes del modelo. Y la probabilidad por si sola sea usada para llevar dicho caso individual a una de las dos categorías de la variable dependiente. (Silva Ayçague, 1995)

Curva Operativa de Rendimiento (ROC)

Con la tabla de clasificación de aciertos, se ha descrito su significado de los porcentajes encontrados para las dos categorías de la variable predicha, entonces todo parte del conocimiento que el punto de corte en el análisis logit es 0.5, porque así lo define por defecto el programa estadístico y la naturaleza de las investigaciones lo disponen de forma general de esa manera. En última parte de la regresión logit se consigue la curva ROC, (por sus siglas en inglés) cuyos ejes son la sensibilidad o susceptibilidad (eje y) y el complemento de la especificidad (eje x) y al formar las coordenadas para la curva se obtiene los valores de referencia para el patrón de desempeño de las proporciones, y ver la relación que guarda con la variable de estado.

La utilidad inicial de esta curva es observar los balances entre sus dos ejes, es probable que se esté interesado en modificar el punto de corte debido a la experiencia de los investigadores y la investigación comience nuevamente con una percepción diferente, pero hay que tener en cuenta que esto genera un análisis previo de la distribución inicial de las categorías de estudio. (Silva Ayçague, 1995)

CAP. III METODOLOGIA

3.1. Diseño de la Investigación

El tipo de investigación implementada tiene las siguientes características:

- No experimental, porque no se manipula los datos extraídos de la población que será objeto de nuestra investigación.
- Retrospectiva, porque se utilizarán datos cuyas fuentes provienen de la empresa CGT.
- Transversal, porque se realizará una sola medición de los factores durante el periodo de duración de la investigación.
- Analítica, porque se va a analizar un conjunto de variables independientes que nos posibilitara estimar el incumplimiento de pago de los contribuyentes.
- Predictiva, porque el modelo obtenido nos va a permitir predecir el incumplimiento de pago de los contribuyentes.

H_o = Los factores (X_1, X_2, X_3,…, X_{12}) no generan un modelo correctamente ajustado a la probabilidad de ser un contribuyente incumplido durante una Campaña de amnistía.

H_1 = Los factores (X_1, X_2, X_3,…, X_{12}) generan un modelo correctamente ajustado a la probabilidad de ser un contribuyente incumplido durante una Campaña de amnistía.

.

3.2. Población y muestra

Para nuestra población se tomó como referencia a todos aquellos contribuyentes de la Municipalidad Provincial de Chiclayo con deuda tributaria (impuesto predial y arbitrios)

pendiente quienes son el público objetivo de las campañas de amnistía tributaria. Siendo un total de 56369 contribuyentes.

Al tener una población cuyos elementos comparten la misma propiedad de interés que el conjunto general de datos y determinar que existe la misma probabilidad de seleccionar cualquier elemento en particular, se tuvo dos opciones dado a que se tiene la Base de Datos de contribuyentes.

Trabajar con toda la población o con una muestra representativa. Tomando en cuenta las restricciones de completitud e inconsistencia de un buen porcentaje de los datos se determinó trabajar con una muestra, para lo cual se utilizó la fórmula que se orienta sobre el cálculo de muestra para datos globales.

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

En donde:

n = es el tamaño de la muestra poblacional a obtener.

N = 56369

σ = 0.5

Z = El grado de confianza utilizado es 95% (1.96)

e = 2%

Por lo tanto, el tamaño de la muestra total es de 2303 elementos, de los cuales 1642 contribuyentes participaron en la campaña de amnistía.

De lo anterior obtuvimos que nuestra muestra efectiva para la obtención del modelo de predicción se trabajó con los 1642 contribuyentes que fueron de alguna u otra forma convocados a participar de la Campaña de Amnistía Tributaria del año 2017.

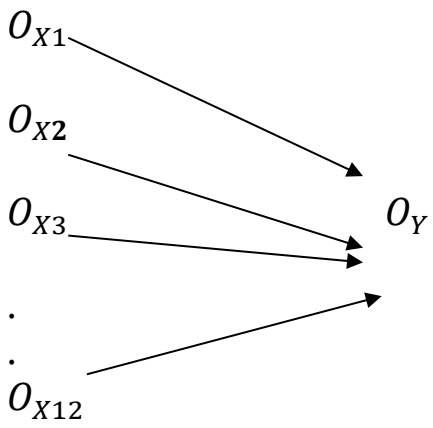
Cuadro 3
Composición de la muestra del modelo

Cartera de Contribuyentes MPCH – Periodo 2017			
Pago Parcial	**Pago Total**	**Impagos**	**Total**
87	164	1391	1642
Muestra de Estimación		50%	821
Muestra de Validación		50%	821

FUENTE: Elaboración Propia

3.3. Diseño de la Investigación

El diseño del presente trabajo de investigación tiene el esquema siguiente:



Modelo de predicción de incumplimiento de pago durante una campaña de amnistía
Fuente: Elaboración propia

Dónde:

O_X: Observación de los factores sociales, económicos, geográficos y demográficos.

O_y : Observación de incumplimiento del pago de arbitrios.

3.4. Operacionalización de las variables

Cuadro 4
Operacionalización de variables

VARIABLES INDEPENDIENTES	INDICADOR	DIMENSIONES	TIPO DE VARIABLE
Número de predios	Número de predios	Unidad	Numérica, discreta
Autovaluo	Valor del terreno, valor de construcciones, valor de otras instalaciones	Soles	Numérica, continua
Tipo de domicilio	Escala definida por investigador	Cercado de Chiclayo, Urbanización, Pueblo joven	Categórica, nominal
Tipo de contribuyente	Escala definida por CGT	Persona natural, Sociedad conyugal, Sucesión indivisa, Persona jurídica	Categórica, nominal
Calificación contributiva	Escala definida por CGT por tramos de autovaluo	Prico, Meco, Peco	Categórica, nominal
Transferencia de venta de predios	Número de predios	Unidad	Numérica, discreta
Genero	Sexo	Masculino, Femenino	Categórica, nominal
Registro de Correo electrónico	Registro en Base de Datos	Tiene registro, no tiene registro	Categórica, nominal
Registro telefónico	Registro en Base de Datos	Tiene registro, no tiene registro	Categórica, nominal
Fiscalización de predio	Registro en Base de Datos	Tiene registro, no tiene registro	Categórica, nominal
Notificación de amnistía	Registro en Base de Datos	Tiene registro, no tiene registro	Categórica, nominal
Segmentación por tipo de pago	Categorías definidas por clustering	Segmento 1, Segmento 2, Segmento 3	Categórica, nominal
VARIABLE DEPENDIENTE	**INDICADOR**	**DIMENSIONES**	**TIPO DE VARIABLE**
Incumplimiento de Pago	Saldo de Deuda Vencida, Pagos efectuados	No pago, Pago Parcial, Pago Total	Categórica, nominal

FUENTE: Elaboración propia

CAP. IV RESULTADOS

4.1. Análisis de los Resultados

4.1.1. Construcción del Modelo Logit

Especificación del Modelo de Regresión Logística

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}}$$

Hipótesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \cdots \beta_{12} = 0$$

$$H_1 : Al\ menos\ un\ \beta_i \neq 0\ \forall i = 1,2,3,4,5,6, \ldots ,12$$

Ho: Los factores (X_1, X_2, X_3,…, X_{12}) no influyen significativamente en el modelo.

H_1: Los factores (X_1, X_2, X_3,…, X_{12}) influyen significativamente en el modelo.

4.1.2. Ajuste del Modelo de Regresión Logística

El ajuste ha sido efectuado mediante el Software de Minería de datos SPSS Modeler 18.0; A partir de este punto se han incluido todos los cuadros necesarios, que tienen directa participación en el ajuste del modelo que se ha diseñado, puesto que se evalúa una variable dependiente o de respuesta y se quiere encontrar un modelo estadístico que la estime a través de su expresión matemática, construida a partir de las variables independientes o influyentes que participan en el modelo.

Cuadro 5
Resumen de procesamiento de casos

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	1642	100%
	Casos perdidos	0	0%
	Total	1642	100%
Casos no seleccionados		0	0%
Total		1642	100%

FUENTE: Elaboración propia

Los cuadros de resumen del procesamiento de los casos nos muestran de manera general un resultado importante previo antes de realizar el análisis, el cuadro presenta que del total de la muestra seleccionada y procesada ningún dato ha sido excluido. Por lo tanto, se debe continuar con el análisis de resultados, al tener conocimiento que no hay casos perdidos, y no presentara problemas con respecto a este supuesto en la regresión Logit.

Cuadro 6
Codificación de la variable dependiente "Incumplimiento de pago"

Valor original	Valor interno
No Pago	1
Pago Parcial	2
Si Pago	3

FUENTE: Elaboración propia

La variable dependiente se ha codificado en el software SPSS Modeler en tres categorías: No Pago (1), Pago Parcial (2) y Pago Total (3). Conociendo que cada unidad de análisis de nuestra muestra está referida a un contribuyente; La codificación nos permite de una manera integral categorizar la información obtenida de los 1642 contribuyentes, respecto a la categoría de incumplimiento de pago en el que se encuentran cada uno de estos con la finalidad de ejecutar la técnica de minería de datos siguiendo los procedimientos técnicos que exige el modelo logit y así poder realizar sencillas interpretaciones a partir de los resultados.

Bloque Inicial

IBM SPSS Modeler proporciona algunos resultados básicos bajo el título de "Resumen de Procesamiento de Casos". Estos se basan en un modelo logístico que contiene sólo un intercepto (constante). Aunque este modelo no es interesante, si muestra alguna información básica. Primero en el cuadro de resumen de procesamiento de casos (Cuadro 7), indica que el modelo siempre predice la categoría más común (No Pago: Incumplimiento en el pago de arbitrios) y esto es cierto en el 84.7% del total de casos seleccionados inicialmente, este resultado representa el porcentaje base que deberíamos estar dispuestos a conseguir como mínimo en cuanto a la capacidad de predicción del modelo final, y que se explica más adelante con la inclusión de las variables significativas y las predicciones realizadas con el modelo matemático conseguido.

Cuadro 7
Resumen de procesamiento de casos

		N	Porcentaje marginal
Incumplimiento de pago	1.0	1391	84.7%
	2.0	87	5.3%
	3.0	164	10.0%
Validos		1642	100.0%
Perdidos		0	
Total		1642	

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

Bloque 1: Método - Por pasos hacia adelante (Razón de Verosimilitudes)

El modelo obtenido con la configuración del método "Por pasos hacia adelante" del algoritmo de regresión logística, encuentra un subgrupo de variables que maximizan la verosimilitud.

A continuación, se presenta un cuadro en el que se incluyen los pasos para una mejor comprensión. Además, se observa que el análisis por pasos con la selección hacia adelante y el estadístico de razón de verosimilitud, ha tomado tres pasos para conseguir el modelo estadístico. De lo anterior podemos deducir que se seleccionaron tres variables para el modelo y estas serán incluidas en las ecuaciones matemáticas finales. El cuadro de "Resumen de los pasos" muestra el estadístico de significancia del predictor que ingreso en cada uno de ellos, en tanto que el cuadro de "La Información de ajuste de los modelos" nos permite observar una prueba de los coeficientes del modelo final. Así como no es de asombrar que los coeficientes en el último paso sean significativos como en el modelo final. Estos cuadros son necesarios para entender en parte las interpretaciones siguientes que están relacionadas con la función de verosimilitud del modelo, donde el resultado que se muestra se utiliza como valor de comparación entre el modelo de regresión logística inicial y final para este conjunto de datos.

Cuadro 8
Resumen de los pasos

Modelo		Acción	Efecto(s)	Criterios de ajuste de modelo			Pruebas de selección de efecto		
				AIC	BIC	Logaritmo de la verosimilitud -2	Chi-cuadrado[a,b]	gl	Sig.
Paso 0	0	Especificado	Intersección	1560.472	1571.280	1556.472	.		
Paso 1	1	Especificado	Segmentación	549.303	581.725	537.303	1266.897	4	.000
Paso 2	2	Especificado	CalifContrib	541.167	595.204	521.167	15.729	4	.003
Paso 3	3	Especificado	Teléfono	538.883	603.727	514.883	6.146	2	.046

Método por pasos: Avanzar por pasos

a. El chi-cuadrado para la entrada se basa en la prueba de puntuación.

b. El chi-cuadrado para la eliminación se basa en la prueba de razón de verosimilitud.

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

Cuadro 9
Información de ajuste de los modelos

Modelo	Criterios de ajuste de modelo			Pruebas de la razón de verosimilitud		
	AIC	BIC	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	1560.472	1571.280	1556.472			
Final	538.883	603.727	514.883	1041.589	10	.s 000

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

La probabilidad de los resultados observados, dados los cálculos del parámetro, se conoce como verosimilitud. Por lo general se utiliza -2 veces el logaritmo natural de la verosimilitud (-2LL) como una medida del ajuste del modelo, dado que tiene vínculos con la distribución de Chi cuadrado. Un buen modelo que tiene una elevada verosimilitud se traduce en un valor pequeño de -2LL. En un ajuste perfecto -2LL es igual a 0.

En el cuadro de N° 9 se observa que en el Modelo de Sólo Intersección el valor de -2LL es de 1556.472, este valor comparado con el obtenido en Modelo Final que es 514.883, se puede efectivamente determinar que se reduce aproximadamente 1041 unidades. Lo cual quiere decir que conforme ingresaron variables adicionales en el modelo, la bondad de ajuste del mismo empieza a ser mejor lo cual se puede notar en el estadístico de verosimilitud -2LL el cual disminuye. Esta prueba es determinante para establecer un modelo.

En el cuadro anterior presentado lo que se muestra y se resume es el estadístico Chi-cuadrado del modelo que es una prueba estadística de la hipótesis nula, acerca de que los coeficientes para todos los términos del modelo son iguales a cero, que en la relación principal con los objetivos de la investigación viene a ser la hipótesis estadística establecida en la parte inicial

de la justificación de resultados. Lo que se busca dentro de esta prueba ómnibus es verificar si el modelo es adecuado y esto se podrá corroborar con la significancia que se puede obtener al realizar la prueba. El valor del estadístico que obtenemos es igual 1041.589 que viene a ser la diferencia entre el-2LL inicial (un modelo solo con la constante) y el mismo coeficiente -2LL final (el modelo que incluye a todas las variables independientes). Tiene 10 grados de libertad, que representan la diferencia entre el número de parámetros en los dos modelos. Rechazamos la hipótesis nula porque la significancia es muy baja (0.000) y concluimos que el grupo de variables mejora la predicción del logaritmo natural de las oportunidades.

Cuadro 10

Pseudo R cuadrado	
Cox y Snell	0.470
Nagelkerke	0.722

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

También se muestran los valores que son análogos a la R cuadrado en la regresión estándar, pero dada la relación funcional que existe entre la media y la desviación estándar de la variable dependiente en el modelo Logit y por ser una variable nominal de tres categorías, la cantidad de varianza explicada por el modelo se debe definir diferente, la R cuadrado de Cox y Snell es igual 0.470 y la R cuadrado de Nagelkerke es igual a 0.722 de las dos, por lo general se prefiere esta última sobre la primera porque puede llegar a tomar un valor máximo de l. A través de estos valores se puede observar que el modelo solo explica una cantidad baja de la varianza total, aunque en el segundo caso se pueda afirmar que el modelo y sus variables explican en aproximadamente un 72% de la dispersión total que existe en los errores del modelo para este conjunto de datos.

En el método por pasos los estadísticos de determinación suelen aumentar según el número de pasos que se van incluyendo en el análisis. Con los resultados mostrados se puede observar que las variables incluidas no son determinantes de forma conjunta para explicar la variación o dispersión de los errores, y se necesita incluir nuevas variables o factores no considerados en este análisis que ayuden a entender de manera contundente la varianza total, esto se puede definir en la misma forma de interpretación como se hace en el análisis de regresión lineal.

Cuadro 11
Estimaciones de parámetro

Incumplimiento de Pago[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
2.0	Intersección	0.309	0.424	0.529	1	0.467			
	[CalifContrib=1,000]	0.268	0.586	,210	1	0.647	1.308	0.415	4.125
	[CalifContrib=2,000]	-0.198	0.375	,280	1	0.597	0.820	0.393	1.710
	[CalifContrib=3,000]	0[b]	.	.	0	.	.	.	.
	[Teléfono=,000]	0.752	0.373	4,073	1	0.044	2.121	1.022	4.404
	[Teléfono=1,000]	0[b]	.	.	0	.	.	.	.
	[Segmentación=clúster-1]	-5.372	0.468	131,817	1	0.000	0.005	0.002	0.012
	[Segmentación =clúster-2]	-4.528	0.409	122,717	1	0.000	0.011	0.005	0.024
	[Segmentación =clúster-3]	0[b]	.	.	0	.	.	.	.
3.0	Intersección	1.816	0.416	19,062	1	0.000			
	[CalifContrib=1,000]	-1.071	0.608	3,101	1	0.078	0.343	0.104	1.129
	[CalifContrib=2,000]	-1.046	0.396	6,954	1	0.008	0.352	0.162	0.765
	[CalifContrib=3,000]	0[b]	.	.	0	.	.	.	.
	[Teléfono=,000]	0.872	0.369	5,592	1	0.018	2.391	1.161	4.926
	[Teléfono=1,000]	0[b]	.	.	0	.	.	.	.
	[Segmentación =clúster-1]	-24.248	1985.003	,000	1	0.990	2.945E-11	0.000	.[c]
	[Segmentación =clúster-2]	-24.203	2239.162	,000	1	0.991	3.082E-11	0.000	.[c]
	[Segmentación =clúster-3]	0[b]	.	.	0	.	.	.	.

a. La categoría de referencia es: 1.0.

b. Este parámetro está establecido en cero porque es redundante.

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

En el cuadro 11 se muestra el valor de los coeficientes del modelo, obtenidos en el tercer paso de este método, el cual quedara como modelo final de presentación, esta última salida debe verse tan igual como la interpretación de la regresión lineal, también se presentan aparte de la columna β y los errores estándar de los β, los valores de una prueba basada en el estadístico de Wald y su nivel de significancia, como también la columna en la que se presentan los Odds Ratio de cada variable, $Exp(\beta)$. Finalmente se puede observar que el modelo se ha ajustado con un total de 3 variables de las 12 iníciales y que estas son significativas al nivel del 5%.

Para la interpretación correspondiente se debe recordar que el modelo original se encuentra en términos del logaritmo natural de las oportunidades o Logit. Por lo consiguiente, el coeficiente β es el efecto de una unidad de cambio en una variable independiente sobre el logaritmo natural de las oportunidades. El significado se da según las categorías que presente la variable independiente especifica de estudio.

Se debe tomar en cuenta para nuestra interpretación que en el paso previo según El cuadro de Pseudo R Cuadrado, el modelo estadístico encontrado no tiene un buen ajuste, por tanto, se presume que las variables seleccionadas no serán las adecuadas para el modelo, por lo cual con fines académicos se hace la visualización de las variables significativas encontradas para obtener una explicación y comprensión conceptual del modelo.

Las variables que disminuyen la oportunidad de caer en el incumplimiento en el pago de los arbitrios (No Pago), según la descripción anterior es solo una (1) denominada “Teléfono” que identifica si tenemos registrado o no el número telefónico del contribuyente, presentando esta variable un efecto positivo en el cuadro mostrado. Mientras que por otro lado las variables que incrementan la oportunidad de caer en el incumplimiento en el pago de los arbitrios (No Pago) son dos (2) identificadas como “CalifContrib” y “Segmentación” estas variables presentan un efecto negativo según el cuadro mostrado.

Aunque se conoce que el coeficiente carece de una interpretación directa sobre su valor, se puede ver el odds ratio donde en el caso de la variable con coeficiente positivo se tiene un valor superior a cero y en el caso de las variables con signo negativo los odds ratio respectivos tienen como resultado valores decimales inferiores a cero, que representan el factor de oportunidad o razón de cambio sobre la variable dependiente al analizar las categorías de la variable independiente, en el cuadro asimismo se presentan los valores de sus respectivos intervalos de confianza y se puede notar que tienen la mismas características numéricas.

Este último valor en paralelo con su intervalo de confianza nos brinda la interpretación más importante para el modelo resultante y cada una de sus variables significativas evaluadas individualmente, sabemos que este resultado se interpreta como un factor sobre la oportunidad en la variable dependiente cuando se realiza un cambio en las categorías de la variable independiente, se conoce también que siempre el modelo Logit hace esto real en términos de la probabilidad que es lo que nos preocupa encontrar de forma más intuitiva y se deduce en la columna $Exp(\beta)$ la interpretación de cada una de las variables incluidas en el modelo que tienen que ver estrictamente con los objetivos que la investigación busca alcanzar dentro del estudio de la institución recaudadora de los tributos municipales, para la comprensión clara de esta situación, se trata de concluir más adelante de forma precisa la interpretación individual de cada una de las variables en el respectivo modelo.

4.1.3. Modelo Final y Odd Ratios

De acuerdo a los resultados del modelado con regresión logística en SPSS mediante el método "Adelante" y con el criterio de Razón de Verosimilitudes se ha obtenido el siguiente modelo, y puesto que la investigación comprende a una variable dependiente llamada Incumplimiento de Pago, se expresaran las ecuaciones finales como sigue a continuación, teniendo una forma

general para el modelo encontrado con la especificación de cada una de las variables significativas que participan de la ecuación matemática y que se expresan continuación:

Ecuación 2 = 0.2685 * [CalifContrib=1] + -0.1984 * [CalifContrib=2] + 0.7521 * [Teléfono=0] + -5.372 * [Segmento= 1] + -4.528 * [Segmento= 2] + + 0.3087

Ecuación 3= -1.071 * [CalifContrib=1] + -1.046 * [CalifContrib=2] + 0.8719 * [Teléfono=0] + -24.25 * [Segmento= 1] + -24.2 * [Segmento= 2] + 1.816

El modelo si bien no se ajusta de acuerdo a la prueba de Bondad de Cox-Snell y Nagelkerke, no obstante, muestra la existencia de tres variables significativas. Con el modelo obtenido se debe corroborar la capacidad predictiva correcta que se logre y se pueda establecer una conclusión final, aunque de antemano se conoce que un ajuste adecuado del modelo es suficiente para considerar el análisis estadístico como apropiado, por lo menos con este tipo de información y con estas variables debemos tener en cuenta que la respuesta de nuestro objetivo de investigación no es favorable, aunque de forma aplicativa y practica se ha alcanzado un ejemplo de uso de la técnica estadística y de los elementos que se deben considerar para realizar un análisis de esta naturaleza.

En lo que respecta al cálculo de los odds ratio de cada factor significativo debemos tener en cuenta que ya se consideró su explicación en la parte anterior a estos resultados, el odds ratio se encuentra al hallar la siguiente expresión $Exp(\beta)$ teniendo en cuenta que solo nos interesa tener el valor de esta expresión matemática; Para los factores considerados en esta investigación que resultaron significativos, como se ha establecido y se puede observar en la cuadro de variables incluidas en los modelos de regresión logística, dicha operación matemática nos ofrece un resultado mayor a cero el cual debe ser interpretado de acuerdo a las categorías de la variable al cual le corresponde en comparación o relación con la variable dependiente de esta investigación.

Para el incumplimiento de Pago los factores que resultaron significativos fueron los siguientes con su respectivo odds ratio calculado a partir del valor del coeficiente estimado para las ecuaciones resultantes:

Cuadro 12
Resumen de los parámetros estimados

Variables Predictoras		**Pago Parcial**			**Pago Total**		
		$\boldsymbol{\beta}$	$\boldsymbol{Exp(\beta)}$	**Sig.**	$\boldsymbol{\beta}$	$\boldsymbol{Exp(\beta)}$	**Sig.**
CalifContrib	**Prico**	0.268	1.308	0.647	-1.071	0.343	0.078
	Meco	-0.198	0.820	0.597	-1.046	0.352	0.008
Registro Telefónico	**No**	0.752	2.121	0.044	0.872	2.391	0.018
	Si	0	.	.	0	.	.
Segmentación	**Cluster-1**	-5.372	0.005	0	-24.248	0	0.990
	Cluster-2	-4.528	0.011	0	-24.203	0	0.991

FUENTE: Elaboración propia

Del cuadro anterior, y tomando en consideración la Significancia al 0.05% de los respectivos Odd Ratios determinados, debemos concluir con respecto a aquellos contribuyentes que efectuaron un Pago Parcial:

- Que aquellos contribuyentes que tienen el “Registro Telefónico” igual a No, tienen 2.121 más posibilidad de efectuar el Pago Parcial de su deuda respecto al No Pago de la misma en la próxima campaña de amnistía tributaria.
- Que aquellos contribuyentes que tienen la “Segmentación” igual a Clúster-1 y Clúster-2, definitivamente tienen una nula posibilidad de efectuar el Pago Parcial de la misma en la próxima campaña de amnistía tributaria.

Así mismo del cuadro anterior, y tomando en consideración la Significancia al 0.05% de los respectivos Odd Ratios determinados, debemos concluir con respecto a aquellos contribuyentes que efectuaron un Pago Total:

- Que aquellos contribuyentes que tienen el "Registro Telefónico" igual a No, tienen 2.391 más posibilidad de efectuar el Pago Total de su deuda respecto al No Pago de la misma en la próxima campaña de amnistía tributaria.
- Que aquellos contribuyentes que tienen el "CalifContrib" igual a Meco, tienen 0.352 menos posibilidad de efectuar el Pago Total de su deuda respecto al No Pago de la misma en la próxima campaña de amnistía tributaria.

4.1.4. Validación del Modelo

Cuadro 13
Clasificación de los resultados de la predicción

Clasificación				
	Pronosticado			
Observado	**1.0**	**2.0**	**3.0**	**Porcentaje correcto**
1.0	1358	0	33	97.6%
2.0	25	0	62	0.0%
3.0	0	0	164	100.0%
Porcentaje global	84.2%	0.0%	15.8%	92.7%

FUENTE: Modelo de Regresión Logística. SPSS Modeler. Elaboración propia

Para la validación del presente modelo, se separó una muestra (de validación) de 821 contribuyentes. Si aplicamos las ecuaciones logísticas estimadas a la muestra de validación el porcentaje global de acierto es del 92.7% como se muestra en el cuadro 12. Es así que se llega a predecir en un 97.6 % a los contribuyentes incumplidos (No Pago= 1), en un 0% a los

contribuyentes incumplidos (Pago Parcial = 2) y en 100% a los contribuyentes incumplidos (Pago Total= 3).

4.2. Discusión de los Resultados

4.2.1. Hipótesis General

Se determino el contraste de la hipótesis considerando el alcance predictivo del modelo, por lo que se efectuó la validación del mismo a través de una serie de pruebas que miden distintos aspectos del modelo, los que permitió determinar la aprobación de la hipótesis nula.

Ho: Los factores (X_1, X_2, X_3,…, X_{12}) no generan un modelo correctamente ajustado a la probabilidad de ser un contribuyente incumplido en el pago de arbitrios durante la Campaña de Amnistía Tributaria del periodo 2018

Cuadro 14
Contrastación de la Hipótesis General

VARIABLE	HIPOTESIS	INDICES	INTERPRETACIÓN	DIMENSIONES
Incumplimiento de Pago	Los factores X1, X2, X3,…, X12 no generan un modelo correctamente ajustado a la probabilidad de ser un contribuyente incumplido en el pago de arbitrios durante una	- Razón de verosimilitud(-2LL) - Ajuste del modelo (Pseudo R cuadrado de Cox-Snell y Nagelkerke)	-2LL es 1041.589, es mucho mayor que CERO, lo cual no es una razón adecuada El modelo explica entre 47% y 72.2% la dispersión de sus errores. Lo cual es un valor relativamente bajo Precisión es de 92.7%, la cual es Alta	Solo tres variables independientes contribuyen al modelo.

	Campaña de Amnistía Tributaria del periodo 2018	- Precisión (de la predicción) - Odd Ratios		

FUENTE: Elaboración propia

Los porcentajes globales de predicciones correctas, es un resultado que nos da a conocer que tan bueno es el modelo en función a las predicciones que llegue a realizar en el futuro, en este caso el modelo supera el 90% lo cual se considera un buen alcance predictivo, no obstante esta apreciación debemos tomarla con prudencia debido a que los casos que están en la categoría de interés de la variable dependiente son la mayoría de casos participes de la muestra, además está la consideración que el modelo no presenta un ajuste adecuado lo que genera controversias en la decisión final que se debe adoptar.

Cuadro 15
Contrastación de Hipótesis H1 por Dimensiones

DIMENSION	**HIPOTESIS**	**COEFICIENTES**	**MODELO DE INCUMPLIMIENTO DE PAGO**
Tipo De Contribuyente	Contribuye al modelo	No presenta	No interviene en modelo
Calificación Contributiva	Contribuye al modelo	> 0 y < 0	Incrementa y Disminuye probabilidad de incumplimiento
Genero	Contribuye al modelo	No presenta	No interviene en modelo
Número de Predios de su	Contribuye al	No presenta	No interviene en modelo

propiedad	modelo		
Autovaluo afecto	Contribuye al modelo	No presenta	No interviene en modelo
Transferencia de Venta	Contribuye al modelo	No presenta	No interviene en modelo
Tipo de Contribuyente para notificación	Contribuye al modelo	No presenta	No interviene en modelo
Registro de Email	Contribuye al modelo	No presenta	No interviene en modelo
Registro Telefónico	Contribuye al modelo	> 0	Disminuye probabilidad de incumplimiento
Fiscalización de predios	Contribuye al modelo	No presenta	No interviene en modelo
Notificación de Beneficios Tributarios	Contribuye al modelo	No presenta	No interviene en modelo
Segmentación por Comportamiento de Pago	Contribuye al modelo	< 0	Incrementa probabilidad de incumplimiento

Fuente: Elaboración Propia

CONCLUSIONES

La presente investigación permitió concluir que:

A. De las doce variables independientes que se consideraron como entradas para el modelo, solo tres de ellas resultaron ser significativos con una confianza al 95%, razón por la cual se debe concluir que este conjunto de variables parece no ser las más adecuadas para llevar adelante este modelo de predicción, por lo tanto, se deben estudiar nuevas características que aporten y sean más relevantes para la predicción del incumplimiento de pago.

B. Se logro calcular la probabilidad de incumplimiento de pago de cada contribuyente de la muestra de validación, logrando alcanzar una alta precisión de la misma. No obstante, es necesario que se considere la alta predisposición que se tiene al "No Pago" en el total de los elementos de la muestra.

C. Se logro un nivel de precisión alto en el pronóstico del incumplimiento de pago, debido en su conjunto a la sensibilidad del modelo, sin embargo, hay que señalar se debe considerar mejorar la especificidad con el fin de conseguir un mejor modelo en su conjunto.

D. El modelo no logra cumplir con el criterio de controversia, dado a que el modelo logra alcanzar una buena calidad predictiva superior al 92%, pero no logra tener un buen ajuste estadístico.

E. Respecto a la muestra utilizada se observa que no existen los suficientes casos posibles en cada categoría de la variable dependiente, para que el análisis estadístico correspondiente se realice de la mejor forma, en nuestro estudio hay una notoria diferencia entre las categorías No Pago, Pago Parcial y Pago Total.

F. El modelo obtenido aún no puede ser utilizado para el diagnóstico de contribuyentes debido a las dificultades técnicas presentadas.

RECOMENDACIONES

Con los resultados de la presente investigación se recomienda.

A. Ampliar el espectro de variables independientes con el fin de incrementar la posibilidad de encontrar las que contribuyan con el ajuste del modelo y se cierre la brecha aún existente.

B. Equilibrar la muestra considerando los porcentajes de incidencia de cada uno de las categorías de la variable dependiente, con la finalidad de obtener un modelo que discrimine mejor la variabilidad del comportamiento de pago de los contribuyentes.

C. Dado a que el modelo aún no está listo para ser utilizado para la predicción, se recomienda mejorarlo en una próxima etapa con el fin de lograr un análisis más minucioso en el que se considere todos los supuestos a cumplir para el modelo de regresión logística.

.

REFERENCIAS BIBLIOGRÁFICAS

C. d. (2004). *Ley de Tributos Municipales.*

Calixto Salazar, M. M., & Casaverde Carranza, L. F. (2011). *Variables determinantes de la probabilidad de incumplimiento de un microcrédito en una entidad microfinanciera del Perú, una aproximación bajo el modelo de regresión logística binaria.*

Cervantes Canales, J. (2009). *Clasificación de grandes conjuntos de datos vía máquinas de vectores de soporte y aplicaciones en sistemas biológicos* .

Chunga Chully, L. M., & Reyes Pintado, J. F. (2015). *Modelo Logit para determinar los factores socioeconomicos que influyen en el incumplimiento de pago del impuesto predial del distrito de Piura.*

Jélvez Camaño, A., Moreno Echevarria, M., Ovalle Retamal, V., Torres Navarro, C., & Troncoso Espinoza, F. (2014). *Modelo Predictivo de fuga de clientes utilizando míneria de datos para una empresa de telecomunicaciones en Chile.* Paris: Pearson Education France.

Marín Diazaraque, J. M. (2011). *halweb.uc3m.es.* Recuperado el 10 de 11 de 2018, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema5Cate.pdf

Miranda, J., Rey, P., & Weber, R. (2005). *Predicción de fuga de clientes para una institución financiera mediante Support Vector Machine.*

Rayo Cantón, S., Lara Rubio, J., & Camino Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de.

Sampaio Dória, A. (1971). *V Asamblea General del CIAT. Río de Janeiro.* Río de Janeiro.

Silva Ayçague, L. C. (1995). *Excursión a la regresión logística en ciencias de la salud.* España: Ediciones Diaz Santos, S.A.

Vilcapoma, E. (2003). *Identificación de patrones de evasión en el sistema de administración tributaria usando tecnología Datamining.*

ANEXOS

ENTREVISTAS

Entrevista 1

Usuario: Jefe de Cobranzas

PREGUNTAS

1. ¿Cuáles son las carteras de contribuyentes que se trabajan en la gestión de cobranza coactiva?
2. ¿Qué criterios utilizan para la definición de estas carteras?
3. ¿Cuáles son las características de una cartera de PRICOS?
4. ¿Cuáles son las características de un contribuyente asignado a la cartera de PRICOS?
5. ¿Cuál/cuales son las características que tiene un PRICO que ustedes consideren tiene una mayor probabilidad de pago?
6. ¿Para un PRICO qué tiene SOLO deuda en cobranza ordinaria, cuál cree que es el criterio (puede ser por tributo) que primara a la hora de decidir la prioridad de la deuda que va a cancelar?
7. ¿Para un PRICO qué tiene deuda en cobranza ordinaria y coactiva, cuál cree que es el criterio que primara a la hora de decidir la prioridad de la deuda que va a cancelar (Entre deuda en ordinaria ó en coactiva)?
8. ¿Cuál/cuales son las características que tiene un PRICO que tiene POCA probabilidad de pago?
9. ¿Cuáles son las características de una cartera de MECOS?
10. ¿Cuáles son las características de un contribuyente asignado a la cartera de MECOS?
11. ¿Cuál/cuales son las características que tiene un MECO que ustedes consideren tiene una mayor probabilidad de pago?
12. ¿Para un MECO qué tiene SOLO deuda en cobranza ordinaria, cuál cree que es el criterio (puede ser por tributo) que primara a la hora de decidir la prioridad de la deuda que va a cancelar?
13. ¿Para un MECO qué tiene deuda en cobranza ordinaria y coactiva, cuál cree que es el criterio que primara a la hora de decidir la prioridad de la deuda que va a cancelar?
14. ¿Cuál/cuales son las características que cumple un MECO que tiene POCA probabilidad de pago?
15. ¿Cuáles son las características de una cartera de PECOS?
16. ¿Cuáles son las características de un contribuyente asignado a la cartera de PECOS?
17. ¿Cuál/cuales son las características que tiene un PECO que ustedes consideren tiene una mayor probabilidad de pago?
18. ¿Para un PECO qué tiene SOLO deuda en cobranza ordinaria, cuál cree que es el criterio que primara a la hora de decidir la prioridad de la deuda que va a cancelar?
19. ¿Para un PECO qué tiene deuda en cobranza ordinaria y coactiva, cuál cree que es el criterio que primara a la hora de decidir la prioridad de la deuda que va a cancelar?
20. ¿Cuál/cuales son las características que cumple un PECO que tiene POCA probabilidad de pago?
21. ¿Qué criterios utilizan para determinar la necesidad de realizar una campaña de amnistía tributaria?

22. ¿Cuáles son los periodos más frecuentes en los que se desarrollan campañas de amnistía tributaria?
23. ¿Qué tipo de beneficios se ofrecen durante una campaña de amnistía tributaria? ¿Por favor, podría detallármelos?
24. ¿Cuál es el procedimiento que se utiliza para determinar las carteras que se trabajaran durante una campaña de amnistía tributaria?
25. ¿Cuáles son las carteras que se trabajan durante una campaña de amnistía tributaria?
26. ¿Cuánto tiempo les conlleva el determinar una cartera a gestionar?
27. ¿Qué tipo de gestión se hace a cada una de estas carteras durante una campaña de amnistía tributaria?
28. ¿En qué consiste el análisis por cartera que se hace al término de una campaña de amnistía tributaria?
29. ¿Cuáles son los resultados esperados por cartera en una campaña de amnistía tributaria?
30. ¿Qué tipo de factores externos consideran a la hora de determinar su cartera (Ejemplos: Época de elecciones, fenómenos naturales, etc.)?
31. ¿De qué forma se aprovecha la información histórica que se tiene actualmente en sus bases de datos?
32. ¿Cuál es la problemática que se presenta al efectuar la cobranza de los tributos municipales (Respecto al Impuesto Predial y a los Arbitrios municipales)?
33. ¿Cuáles son sus indicadores que evalúa para medir los resultados de sus estrategias de cobranza?

Formato A.1: Entrevista a Jefe de Cobranzas

Entrevista 2
Usuario: Jefe de Sistemas
PREGUNTAS

1. ¿Cómo se explota la información histórica que ha generado el uso de sus aplicaciones?
2. ¿Qué tipo de soluciones tecnológicas usted cree pueden usarse para mejorar el proceso de gestión de la cobranza?
3. ¿Qué herramientas utilizan para obtener y analizar los resultados de las estrategias de cobranza implementadas?

Formato A.2: Entrevistas al Jefe de Sistemas

Cuadro 1
Listado de atributos iníciales extraídos de la Base de Datos

Nombre	**Descripción**
CodCont	Tipo de dato: numérico, número entero que identifica a un contribuyente.
TipoCont	Tipo de dato: nominal, representa el tipo de contribuyente, sus valores pueden ser 01-> Persona natural, 02-> Sociedad conyugal , 03-> Sucesión indivisa y 11-> Persona jurídica.
CalifContrib	Tipo de dato: varchar, cuyos valores pueden ser: 001-> Principal contribuyente, 002->Mediano contribuyente, 003-> Pequeño contribuyente
Sexo	Tipo de dato: varchar sus valores pueden ser: M-> masculino, F-> femenino, I->sin sexo. Este campo también posee valores nulos, debido a que su registro no es obligatorio
CodLugar	Tipo de dato: varchar, que consta de 9 dígitos que identifica el lugar del domicilio fiscal del contribuyente
NroPredios	Tipo de dato: numérico, número entero que indica el número de predios que son propiedad del contribuyente
DirFiscal	Tipo de dato: varchar, contiene el domicilio fiscal completo del contribuyente. Está compuesto por: Lugar, Av./Ca., Dpto, numero, etc.
ValuoAfecto	Tipo de dato: numérico, el valor de este campo representa el valor monetario sobre el que se calculara el monto del impuesto predial del contribuyente
NroPrediosVenta	Tipo de dato: numérico, número entero que indica el número de predios propiedad del contribuyente que han sido vendidos a otro contribuyente.
NroPrediosCompra	Tipo de dato: numérico, número entero que indica el número de predios que han sido adquiridos por un

	contribuyente.
Observado	Tipo de dato: bit, tiene valor “1” cuando el contribuyente tiene una observación específicamente por problemas en el domicilio fiscal, tiene valor “0” en el caso contrario.
Email	Tipo de dato: bit, tiene valor “1” cuando el contribuyente tiene como mínimo un email valido, tiene valor “0” en el caso contrario.
Telefono	Tipo de dato: bit, tiene valor “1” cuando el contribuyente tiene como mínimo un número telefónico valido, tiene valor “0” en el caso contrario.
FisLast3years;	Tipo de dato: bit, tiene valor “1” cuando el contribuyente ha tenido como mínimo una fiscalización de sus predios en los últimos tres años, tiene valor “0” en el caso contrario.
NotifBenefAmnis	Tipo de dato: bit, tiene valor “1” cuando el contribuyente le ha sido notificado el beneficio de amnistía tributaria el año 2017, tiene valor “0” en el caso contrario.
IndicadorDeudaAnoActualPredial	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene deuda del impuesto predial en el año actual, tiene valor “0” en el caso contrario
IndicadorDeudaAnosAnteriores Predial	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene deuda del impuesto predial en los cuatro años anteriores al actual, tiene valor “0” en el caso contrario
IndicadorDeudaAnoActualArbitrios	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene deuda de arbitrios municipales en el año actual, tiene valor “0” en el caso contrario
IndicadorDeudaAnosAnteriores Arbitrios	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene deuda de arbitrios municipales en los cuatro años anteriores al actual, tiene valor “0” en el caso contrario

IndicadorPagoPredialActual	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene pagos del impuesto predial en el año actual, tiene valor “0” en el caso contrario
IndicadorPagoArbitriosActual	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene pagos de los arbitrios municipales en el año actual, tiene valor “0” en el caso contrario
IndicadorPagoPredialAnosAnteriores	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene pagos del impuesto predial en los cuatro años anteriores al actual, tiene valor “0” en el caso contrario
IndicadorPagoArbitriosAnosAnteriores	Tipo de dato: bit, Tiene valor “1” cuando el contribuyente tiene pagos de arbitrios municipales en los cuatro años anteriores al actual, tiene valor “0” en el caso contrario
ElementMuestra	Tipo de dato: numérico, este número entero representa si el contribuyente seleccionado forma parte de la muestra

FUENTE: Elaboración propia