Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ingeniería Electrónica y Eléctrica
Unidad de Posgrado

Modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boosting

TESIS

Para optar el Grado Académico de Magíster en Dirección Estratégica de las Telecomunicaciones

AUTOR

Saúl HUAQUIPACO ENCINAS

ASESOR

Mg. Wilbert CHÁVEZ IRAZABAL

Lima, Perú

2022

Referencia bibliográfica

Huaquipaco, S. (2022). *Modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boosting*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Electrónica y Eléctrica, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Saúl Huaquipaco Encinas
Tipo de documento de identidad	DNI
Número de documento de identidad	43237458
URL de ORCID	https://orcid.org/0000-0003-2323-3061
Datos de asesor	
Nombres y apellidos	Wilbert Chávez Irazabal
Tipo de documento de identidad	DNI
Número de documento de identidad	08121733
URL de ORCID	https://orcid.org/0000-0002-7978-7031
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Nicanor Raúl Benites Saravia
Tipo de documento	DNI
Número de documento de identidad	10189914
Miembro del jurado 1	
Nombres y apellidos	Carlos Alberto Moreno Paredes
Tipo de documento	DNI
Número de documento de identidad	01292577
Miembro del jurado 2	
Nombres y apellidos	Carlos Alberto Sotelo López
Tipo de documento	DNI
Número de documento de identidad	07017259
Miembro del jurado 3	
Nombres y apellidos	Arlich Joel Portillo Allende
Tipo de documento	DNI
Número de documento de identidad	10125495

Datos de investigación	
Línea de investigación	C.0.3.3. Desarrollo de modelos y aplicación de las tecnologías de información y comunicaciones
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.056416 Longitud: -77.081153
Año o rango de años en que se realizó la investigación	2016 - 2018
URL de disciplinas OCDE	Telecomunicaciones https://purl.org/pe-repo/ocde/ford#2.02.05

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
(Universidad del Perú, DECANA DE AMÉRICA)
FACULTAD DE INGENIERÍA ELECTRÓNICA Y ELÉCTRICA
UNIDAD DE POSGRADO
Calle Germán Amezaga N.º 375 Lima (Perú)
Telefóno (01) 6197000 Anexo 4204
Correo: postfie@unmsm.edu.pe

«AÑO DEL FORTALECIMIENTO DE LA SOBERANÍA NACIONAL»

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN DIRECCIÓN ESTRATÉGICA DE LAS TELECOMUNICACIONES

Siendo las 12:00 horas del 24 de mayo de 2022, los suscritos miembros del jurado reunidos en el salón de Grados de la Facultad de Ingeniería Electrónica y Eléctrica, el Jurado Examinador presidido por el Dr. Nicanor Raúl Benites Saravia, Dr. Carlos Alberto Moreno Paredes, Mg. Carlos Alberto Sotelo López, Mg. Arlich Joel Portillo Allende y el Mg. Wilbert Chávez Irazabal.

Se reunió para la sustentación oral y pública de la Tesis para optar el Grado Académico de Magíster en Dirección Estratégica de las Telecomunicaciones, que solicitó el alumno **Saúl Huaquipaco Encinas** con código Nº 14197028, el cual procedió hacer la exposición oral y pública de su Tesis Titulada **"MODELO DE CLASIFICACIÓN Y PREDICCIÓN MULTIVARIABLE DE CALIDAD DE SERVICIO MEDIANTE INDICADORES DE ATENCIÓN UTILIZANDO MÉTODOS HÍBRIDOS DE SELECCIÓN DE VARIABLES Y EXTREME GRADIENT BOOSTING"**

Concluida la exposición, el Jurado Examinador procedió a formular las preguntas reglamentarias y, luego de una deliberación en privado, decidió otorgarle la siguiente calificación:

BUENO	16	dieciséis
	NÚMERO	LETRAS

A continuación, el Presidente Jurado recomienda que la Unidad de Posgrado proceda con el trámite correspondiente para que se otorgue el Grado Académico de Magíster en Dirección Estratégica de las Telecomunicaciones al alumno **Saúl Huaquipaco Encinas.**

Siendo las 13:10 Hrs se levantó la Sesión, recibiendo el graduado las felicitaciones de los señores miembros del Jurado y público asistente.

Dr. NICANOR RAÚL BENITES SARAVIA
Presidente

Dr. CARLOS ALBERTO MORENO PAREDES
Miembro

Mg. CARLOS ALBERTO SOTELO LÓPEZ
Miembro

Mg. ARLICH JOEL PORTILLO ALLENDE
Miembro

Mg. WILBERT CHAVEZ IRAZABAL
Asesor

DEDICATORIA

A ADRA y SAHR por estar a mi lado y ser el motivo.

AGRADECIMIENTO

A mi asesor y docentes de la Maestría en Dirección Estratégica de las Telecomunicaciones de la U.N.M.S.M. por guiarme por la senda de la investigación.

Al Dr. José Emmanuel Cruz de la Cruz por su incondicional y desinteresado apoyo.

INDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Uno de los retos más importantes que persiguen las operadoras de telefonía móvil es lograr mejorar la calidad de servicio, pero conocer las necesidades y comportamiento de más de 40 millones de usuarios que hay en el Perú, involucra el tratamiento y procesamiento de una enorme cantidad de datos; en el presente trabajo de investigación se tuvo como objetivo desarrollar un modelo de predicción de calidad de servicio a través de indicadores de atención utilizando métodos híbridos de selección como el Recursive Feature Elimination (RFE), Shrinkaje L1 L2 y extreme gradient boosting (XGBoost), lográndose obtener una Precisión de hasta el 93.75%, una Exactitud de hasta el 88.89% , una Sensibilidad de hasta el 92.86% y un puntaje F1 de hasta 88.31% además de una reducción de variables de hasta el 53.3%.

Palabras clave: Calidad de servicio, Extreme gradient boosting, Machine learning. Modelo de predicción.

SUMMARY

Currently one of the challenges most important pursued by mobile operators is to improve the quality of service, but knowing the needs and behavior of more than 40 million users in Peru involves the treatment and processing of an enormous amount of data; so in the present research work I address as an objective to develop a model of prediction of quality of service through attention indicators using hybrid selection methods such as Recursive Feature Elimination (RFE), Shrinkage L1 L2 and extreme gradient boosting (XGBoost), achieving a Precision of up to 93.75%, an Accuracy of up to 88.89%, a Sensitivity of up to 92.86% and an F1 score of up to 88.31% in addition to a reduction of variables of up to 53.3%.

Keywords: Quality of service, Extreme gradient boosting, Machine learning. Prediction model .

I. INTRODUCCIÓN

1.1 Situación Problemática.

Actualmente la calidad de servicio tiene una gran importancia en los negocios, estamos en un mercado tan competitivo que los clientes siempre están en la búsqueda de lo mejor, en el sector de telefonía móvil tenemos un mercado como varios a actores como ofertantes y en nuestro país la competencia por captar y preservar a los clientes es competitivo en esta coyuntura la calidad de servicio es uno de los valores que cada empresa no puede dejar de lado.

Hay varios aspectos (indicadores) que los clientes toman en cuenta a la hora de elegir un operador móvil y esos mismos aspectos (indicadores) son los que confluyen para que el cliente decida permanecer en determinado operador, por eso es necesario estar midiendo constantemente esos indicadores y sobre todo poder anticipar en comportamiento del consumidor a fin de poder captarlo y fidelizarlo.

A fin de no perder clientes es necesario usar todas las herramientas posibles para elevar nuestra calidad de servicio una de estas herramientas podría ser el machine learning que a través de algoritmos (Huaquipaco et al. 2021) dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones.

1.2 Formulación del Problema

¿Como influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin?

1.2.1 Problemas Específicos.

1. ¿Cómo influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación?
2. ¿Cómo influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin?

1.3 Justificación.

La calidad de servicio es fundamental a la hora de captar y mantener a un cliente, por ello en un mercado tan competitivo y con varios actores, los operadores tienen anticiparse a las necesidades de los clientes logrando con ello una mejor atención y calidad en el servicio que estos brindan, este proyecto propone desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

1.4.2 Objetivos Específicos.

1. Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación.

2. Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

II. MARCO TEÓRICO

2.1 Marco epistemológico de la investigación.

2.2 Antecedentes del Problema

A hybrid Method with TOPSIS and Machine Learning Techniques for Sustainable Development of Green Hotels Considering Online Reviews

Este artículo propone un método híbrido para el análisis de reseñas en línea a través de la toma de decisiones multicriterio, la minería de textos y las técnicas de aprendizaje predictivo para encontrar la importancia relativa de los factores que afectan la toma de decisiones de los viajeros en la selección de hoteles ecológicos con servicios de spa. El método propuesto se desarrolla por primera vez en el contexto del turismo y la hospitalidad a través de esta investigación, especialmente para la segmentación de clientes en hoteles ecológicos a través de las reseñas en línea de los clientes. Utilizamos el mapa de autoorganización (SOM) para el análisis de conglomerados, la técnica de análisis de Dirichlet latente (LDA) para analizar revisiones textuales, la técnica de orden de preferencia por similitud con la solución ideal (TOPSIS) para clasificar las características del hotel y la técnica neuro-difusa para revelar los niveles de satisfacción del cliente. El impacto de los hoteles ecológicos con servicios de spa y no spa en la satisfacción de los viajeros se investiga para cuatro grupos de viajeros: Viajó solo, Viajó con la familia, Viajó en pareja y Viajó con amigos. El método propuesto se evalúa en las opiniones de los viajeros en 152 hoteles en Malasia. Los hallazgos de este estudio proporcionan un método

importante para la toma de decisiones de los viajeros para la selección de hoteles a través del contenido generado por el usuario (UGC) y ayudan a los gerentes de hoteles a mejorar la calidad del servicio y las estrategias de marketing. (Wu et al. 2020).

Electricity Price Forecasting for Cloud Computing Using an Enhanced Machine Learning Model

La computación en la nube se está apoderando rápidamente de la industria de la tecnología de la información porque hace que la computación sea mucho más fácil sin preocuparse por comprar el hardware físico necesario para los cálculos, sino que estos servicios están alojados por empresas que proporcionan los servicios en la nube. Estas compañías contienen una gran cantidad de computadoras y servidores cuya principal fuente de energía es la electricidad, por lo tanto, el diseño y mantenimiento de estas compañías depende de la disponibilidad de un suministro de energía eléctrica estable y barato. Los centros de nubes están hambrientos de energía. Con los recientes picos en los precios de la electricidad, uno de los principales desafíos en el diseño y mantenimiento de dichos centros es minimizar el consumo de electricidad de los centros de datos y ahorrar energía. La colocación eficiente de datos y la programación de nodos para descargar o mover el almacenamiento son algunos de los principales enfoques para resolver estos problemas. En este artículo, proponemos un modelo extreme gradient Boosting (XGBoost) para descargar o mover el almacenamiento, predecir el precio de la electricidad y, como resultado, reducir los costos de consumo de energía en los centros de datos. El rendimiento de este método se evalúa en un conjunto de datos del mundo real proporcionado

por el Operador Independiente del Sistema Eléctrico (IESO) en Ontario, Canadá, para descargar el almacenamiento de datos en los centros de datos y disminuir eficientemente el consumo de energía. Los datos se dividen en un 70% de entrenamiento y un 30% de pruebas. Hemos entrenado nuestro modelo propuesto en los datos y validamos nuestro modelo en los datos de prueba. Los resultados indican que nuestro modelo puede predecir los precios de la electricidad con un error cuadrático medio (MSE) de 15,66 y un error absoluto medio (MAE) del 3,74% respectivamente, lo que puede resultar en una reducción del 25,32% en los costes de electricidad. La precisión de nuestra técnica propuesta es del 91%, mientras que la precisión de los algoritmos de referencia RF y SVR es del 89% y 88%, respectivamente. (Albahli, Shiraz, and Ayub 2020).

Genetic algorithm based optimized feature engineering and hybrid machine learning for effective energy consumption prediction

Las redes inteligentes se están desarrollando rápidamente, lo que lleva a la necesidad de pronósticos precisos del consumo de energía. Sin embargo, desarrollar un modelo preciso de series temporales para la previsión energética es difícil. Tiene que ser entrenado utilizando características meteorológicas óptimas como la temperatura y los retrasos de tiempo para calificar para un modelo beneficioso. Hemos propuesto un enfoque que utiliza un modelo de aprendizaje automático conjunto basado en XGBoost, algoritmos de regresor vectorial de soporte (SVR) y K-nearest neighbors (KNN). También hemos utilizado el algoritmo genético (GA) para predecir el consumo de carga total a partir de la selección óptima de características. El uso de los datos de consumo de electricidad de la isla de Jeju como un estudio de caso muestra que el modelo

de conjunto propuesto optimizado con GA es más preciso que los modelos individuales de aprendizaje automático. Utilizando solo las características meteorológicas y de tiempo mejor seleccionadas, el modelo propuesto registra todas las características de una serie temporal complicada y muestra una reducción en el error porcentual absoluto medio (MAPE) y el error de registro cuadrático medio raíz para los pronósticos de la semana siguiente. Obtuvimos un 3,35% MAPE de los datos de prueba de tres meses aplicando el modelo propuesto. Los operadores de redes inteligentes pueden administrar los recursos de manera efectiva para proporcionar excelentes servicios a los consumidores en función de los resultados del modelo recomendado (Khan and Byun 2020).

ML defense: Against prediction API threats in cloud-based machine learning service

El aprendizaje automático (ML) ha demostrado su impresionante rendimiento en el mundo moderno, y muchas corporaciones aprovechan la técnica del aprendizaje automático para mejorar la calidad de su servicio, por ejemplo, DeepFace de Facebook. Los modelos de aprendizaje automático con una colección de datos privados procesados por un algoritmo de entrenamiento se consideran cada vez más confidenciales. Los modelos confidenciales suelen entrenarse en un servidor en la nube centralizado, pero de acceso público. El sistema ML-as-a-service (MLaaS) es uno de los ejemplos en ejecución, donde a los usuarios se les permite acceder a modelos entrenados y se les cobra sobre una base de pago por consulta. Desafortunadamente, investigadores recientes

han demostrado la tensión entre el acceso público y los modelos confidenciales, donde se abusa del acceso adversario a un modelo para duplicar la funcionalidad del modelo o incluso aprender información confidencial sobre individuos (que se sabe que está en el conjunto de datos de entrenamiento). Concluimos estos ataques como amenazas API de predicción para simplificar. En este trabajo, proponemos ml defense, un marco para defenderse contra las amenazas de API de predicción, que funciona como un complemento de los sistemas MLaaS existentes. Hasta donde sabemos, este es el primer trabajo que propone una contramedida técnica a los ataques superados por accesos excesivos a consultas. Nuestra metodología no modifica ningún clasificador ni degrada la funcionalidad del modelo (por ejemplo, redondea los resultados). El marco consta de uno o más simuladores y un auditor. El simulador aprende el conocimiento oculto de los adversarios. Luego, el auditor detecta si existe una violación de la privacidad. Discutimos las dificultades intrínsecas y declaramos empíricamente la eficiencia y viabilidad de nuestros mecanismos en diferentes modelos y conjuntos de datos. (Hou et al. 2019).

Generating the blood exposome database using a comprehensive text mining and database fusion approach

Este articulo científico estudia la generación de la base de datos de expo somas sanguíneos utilizando un enfoque completo de minería de texto y fusión de bases de datos que integra una amplia gama de algoritmos de machine learning supervisados y no supervisados a mediana escala. Se hace hincapié en la facilidad de uso, el rendimiento, la documentación y la coherencia de la interfaz

de programación de aplicaciones bajo la licencia BSD simplificada, promoviendo su uso en el área académica y comercial (Barupal and Fiehn 2019).

Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions

La mayoría de los estudios sobre modelos de evaluación del riesgo de crédito para instituciones financieras durante los últimos años se centran en la mejora de los datos desequilibrados o en la mejora de la precisión de la clasificación con modelos multietapa. Si bien el modelado multietapa y el preprocesamiento de datos pueden aumentar un poco la precisión, la naturaleza heterogénea de los datos puede afectar la precisión de clasificación de los clasificadores. Este documento pretende utilizar el clasificador, eXtreme gradient boosting tree (XGBoost), para construir un modelo de evaluación de riesgo de crédito para instituciones financieras. El submuestreo basado en clústeres se implementa para procesar datos desequilibrados. Finalmente, el área bajo la curva operativa del receptor y la precisión de las clasificaciones son los indicadores de evaluación, en la comparación con otros clasificadores de una sola etapa de uso frecuente, como la regresión logística, los algoritmos de autoorganización y la máquina de vectores de soporte. Los resultados indican que el clasificador XGBoost utilizado por este trabajo logra mejores resultados que los otros tres y puede servir como una herramienta superior para el desarrollo de modelos de riesgo de crédito para instituciones financieras. (Chang, Chang, and Wu 2018).

XGBoost : eXtreme Gradient Boosting

Este es un documento sobre el uso del paquete XGboost en R. XGboost es la abreviatura de eXtreme Gradient Boosting package. Es una implementación eficiente y escalable del marco de aumento de gradiente por (Friedman, 2001) (Friedman et al., 2000). El paquete incluye un eficiente solucionador de modelos lineales y un algoritmo de aprendizaje de árboles. Soporta varias funciones objetivas, incluyendo regresión, clasificación y clasificación(Chen, He, and Benesty 2018).

Quality Assurance of Machine Learning Software

Las funcionalidades del software de aprendizaje automático dependen de un conjunto de datos que se introduzcan en ellas; un ligero cambio en un conjunto de datos de entrenamiento tiene mucho impacto en los valores de los parámetros de aprendizaje y, por lo tanto, en los resultados de inferencia. Los sistemas basados en ML plantean un nuevo desafío a los métodos de garantía de calidad. Este documento revisa dos puntos de vista tradicionales de las cualidades del servicio y del producto. Además, introduce una visión de plataforma, en la que la co-creación de valor es una preocupación importante. Términos de índice: diversidad de conjuntos de datos, pruebas metamórficas, lógica dominante del servicio, evaluación independiente (Nakajima 2018).

Análisis de Comportamiento de Consumos de Clientes

Esta tesis de posgrado tuvo como objetivo el procesamiento de la data de ventas de los tres últimos años de la marca Café Candelas usando modelos estadísticos y machine learning. Primero reviso luego corrigió y finalmente realizo la normalización de los data base provistas por la marca Café Candelas luego las consolido y archivo; para poder determinar que provincias tenían preferencias similares de consumo. Como método final aplicaron métodos de regresión lineal determinando así la proyección de ventas para los siguientes años (Valenzuela Najar 2018).

UCI Machine Learning Repository

Machine learning de la UCI es una recopilación de bases de datos, teorías de dominio y generadores de datos utilizados por los usuarios e interesados en machine learning para analizar algoritmos, creado como un archivo ftp en 1987 por David Aha y junto a alumnos de posgrado en Irvine. El sitio web actual es del 2007 por Rexa.info junto a Asunción y Newman con de la Universidad de Massachusetts Amherst. Financiado por la Fundación Nacional de Ciencias (Dua, Dheeru and Graff 2017).

Predicción de fuga de clientes: una aplicación de técnicas de data mining en telefonía móvil

Este trabajo analizo la información histórica de consumo, usando la comparación de distintos modelos predictivos como la Máquinas de Vectores Soporte, Regresión Logística, Gradient Boosting, Random Forests, Redes

Neuronales y Árboles de Decisión. Teniendo como variables independientes a, consumo, satisfacción y contacto, obtenidas por seis meses. Y considero como variable dependiente a “baja del cliente”, durante el octavo mes del 2014. La novedad de esta investigación radica en el uso repetición de ensayos usando diferentes semillas y de métodos de validación cruzada, concluyendo que el modelo basado en Gradient Boosting fue el mejor para rastrear clientes con riesgo de fuga (Cajachahua Espinoza 2015).

MACHINE LEARNING APPROACHES IN IMPROVING SERVICE LEVEL AGREEMENT-BASED ADMISSION CONTROL FOR A SOFTWARE-AS-A-SERVICE PROVIDER IN CLOUD

El software como servicio (SaaS) ofrece un acceso confiable a las aplicaciones de software a los usuarios finales a través de Internet sin inversión directa en infraestructura y software. Los proveedores de SaaS utilizan recursos de centros de datos internos o alquilan recursos de un proveedor público de infraestructura como servicio (IaaS) para servir a sus clientes. El alojamiento interno puede tener un amplio costo de administración y mantenimiento, mientras que la contratación de un proveedor de IaaS puede afectar la calidad del servicio debido a su rendimiento variable. Para superar estos inconvenientes, proponemos algoritmos pioneros de control de admisión y programación para que los proveedores de SaaS utilicen de manera efectiva los recursos de la nube pública para maximizar las ganancias al minimizar el costo y mejorar el nivel de satisfacción del cliente. Hay un inconveniente en este método es la fortaleza de los algoritmos al manejar errores en escenarios dinámicos de entorno de nube, también existe la necesidad de un método de aprendizaje automático para

predecir las estrategias y producir los recursos correspondientes. El control de admisión proporcionado por el modelo de confianza que se basa en SLA utiliza diferentes estrategias para decidir sobre la aceptación de las solicitudes de los usuarios de modo que haya un impacto mínimo en el rendimiento, evitando las penalizaciones de SLA que están dando mayores ganancias. El método de aprendizaje automático tiene como objetivo construir un sistema distribuido para el monitoreo y la predicción de recursos en la nube que incluya metodologías basadas en el aprendizaje para el modelado y la optimización de los modelos de predicción de recursos. Los métodos de aprendizaje son Artificial Neural Network (ANN) y Support Vector Machine (SVM) son dos estrategias típicas de aprendizaje automático en la categoría de computación de regresión. Estos dos métodos se pueden emplear para modelar la predicción del estado de los recursos. Además, llevamos a cabo un amplio estudio de evaluación para analizar qué solución coincide mejor en qué escenario para maximizar las ganancias del proveedor de SaaS. Los resultados obtenidos a través de nuestra extensa simulación muestran que nuestros algoritmos propuestos proporcionan una mejora significativa (hasta un 40% de ahorro de costos) sobre los de referencia de la literatura. (Mohana and Thangaraj 2013).

Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms

La optimización secuencial basada en modelos (también conocida como optimización bayesiana) es uno de los métodos más eficientes (evaluación por función) de minimización de funciones. Esta eficiencia lo hace apropiado para optimizar los hiper parámetros de los algoritmos de aprendizaje automático que

tardan en entrenarse. La biblioteca Hyperopt proporciona algoritmos e infraestructura de paralelización para la optimización de hiper parámetros por formación (selección de modelos) en Python. Este artículo presenta un tutorial introductorio sobre el uso de la biblioteca Hyperopt, que incluye la descripción de los espacios de búsqueda, la minimización (en serie y en paralelo) y el análisis de los resultados recopilados en el curso de la minimización. El documento se cierra con un debate sobre la labor en curso y futura. (Bergstra, Yamins, and Cox 2013).

ANÁLISIS SOBRE LA NECESIDAD DE REGULAR LA CALIDAD DEL SERVICIO DE TELEFONÍA MÓVIL EN EL PERÚ

Mellado Ochoa analiza la justificación de la regulación la calidad del servicio de telefonía móvil en el Perú, estudiando las características y competencia del mercado de telefónico móvil, analizando si la calidad de servicio se autorregula por la competencia o por el contrario se degrada, concluyendo que cuando la calidad de servicio era autorregulada esta se degradaba por lo que justificaba su regulación por parte de las autoridades competentes. Finalmente, efectuó recomendaciones al marco regulatorio respecto a la calidad del servicio. (Mellado Ochoa 2013).

Web service quality control based on text mining using support vector machine

Los sitios web populares pueden ver cientos de mensajes publicados por día. Los mensajes clave para el departamento de servicio al cliente son las quejas de los clientes, incluidos los problemas técnicos y los informes no

satisfactorios. En este estudio se propone un mecanismo automático para clasificar los mensajes de los clientes en función de las técnicas de minería de texto y máquina de vectores de soporte (SVM). El mecanismo propuesto puede filtrar los mensajes en las quejas de forma automática y adecuada para mejorar la productividad del departamento de servicio y la satisfacción del cliente. Este estudio emplea el gráfico p-control para controlar la tasa de quejas por debajo del nivel de calidad de servicio esperado para la ejecución del sitio web. Este estudio adopta un sitio web de la comunidad como ejemplo. Los resultados experimentales demostraron que, a saber, la capacidad del SVM para reconocer correctamente los mensajes defectuosos superó el 83% con un promedio del 89% para el mecanismo de clasificación, y el gráfico pcontrol fue capaz de reflejar cambios inusuales en la calidad del servicio oportunamente. (Lo 2008).

Additive Logistic Regression

El impulso es uno de los desarrollos recientes más importantes en la metodología de clasificación. Boosting funciona aplicando secuencialmente un algoritmo de clasificación a versiones reponderadas de los datos de entrenamiento y luego tomando un voto mayoritario ponderado de la secuencia de clasificadores así producida. Para muchos algoritmos de clasificación, esta estrategia simple resulta en mejoras dramáticas en el rendimiento. Mostramos que este fenómeno aparentemente misterioso puede entenderse en términos de principios estadísticos bien conocidos, a saber, el modelado aditivo y la máxima probabilidad. Para el problema de las dos clases, el impulso puede verse como una aproximación al modelado aditivo en la escala logística utilizando la máxima probabilidad de Bernoulli como criterio. Desarrollamos aproximaciones más

directas y mostramos que exhiben resultados casi idénticos a los del impulso. Se derivan generalizaciones multiclase directas basadas en la probabilidad multinomial que exhiben un rendimiento comparable a otras generalizaciones multiclase propuestas recientemente de impulso en la mayoría de las situaciones, y muy superior en algunas. Sugerimos una modificación menor al impulso que puede reducir el cálculo, a menudo por factores de 10 a 50. Finalmente, aplicamos estos conocimientos para producir una formulación alternativa de impulsar los árboles de decisión. Este enfoque, basado en la inducción del árbol truncado de la mejor primera, a menudo conduce a un mejor rendimiento y puede proporcionar descripciones interpretables de la regla de decisión agregada. También es mucho más rápido computacionalmente, lo que lo hace más adecuado para aplicaciones de minería de datos a gran escala. (Friedman, Hastie, and Tibshirani 2000).

2.3 Bases Teóricas.

2.3.1 Machine Learning.

El Machine Learning (Aprendizaje Automático) es un campo de investigación que fusiona la estadística, inteligencia artificial y la informática. La aplicación de métodos de machine learning es cada vez más popular varios sectores de la vida cotidiana como las redes sociales, servicios de streaming, compra de productos, perfiles de gustos musicales, sitios web, etc***.(****Andreas and Sarah 2017****).***

2.3.1.1 Tipos de Machine Learning.

Actualmente el Machine Learning se puede agrupar en 3 tipos:

2.3.1.2 Supervised Learning.

El Supervised Learning (Aprendizaje Supervisado) esta técnica genera información basado en el análisis de datos etiquetados. Es una de las técnicas más utilizadas en el machine learning para diversos problemas como son la detección de correo spam, detectores de captchas, análisis de perfiles, etc. dentro de supervised learning podemos encontrar los métodos de clasificación y regresión.(Andreas and Sarah 2017)

2.3.1.3 Unsupervised Learning.

El Unsupervised Learning (Aprendizaje no Supervisado) esta técnica genera información basado en el análisis de datos sin etiquetar a diferencia del Supervised Learning el aprendizaje analiza la información que aún no tiene resultados. Dentro del Unsupervised Learning tenemos la técnica del clustering y reducción dimensional que son usadas en estrategias de diseño de segmentos de mercado, características comunes es mercados reducción de redundancia de información etc.(Andreas and Sarah 2017)

2.3.1.4 Reinforcement Learning.

El Reinforcement Learning (aprendizaje reforzado) forma parte de lo que hoy se conoce como Deep Learning (aprendizaje profundo) y es un técnica del

machine learning que se centra en mejorar la performance en base al análisis de resultados ya procesados.(Andreas and Sarah 2017)

2.3.2 Python

Actualmente Python es uno de los lenguajes de programación más populares para la aplicación de ciencia de datos. Es de código abierto y se presenta como una alternativa muy robusta a MATLAB o R. tiene bibliotecas para procesar imágenes, lenguaje naturas, datos, etc. (Andreas and Sarah 2017)

2.3.3 eXtreme Gradient Boosting

El aprendizaje automático ganó reconocimiento con la primera red neuronal en los años 1940, seguido por el primer campeón verificador de aprendizaje automático en los años 1950. Después de algunas décadas, el campo del aprendizaje automático despega cuando Deep Blue venció al famoso campeón mundial de ajedrez Gary Kasparov en los años 1990, con un aumento en el poder computacional, los años 1990 y principios de la década de 2000 produjeron una gran cantidad de artículos académicos que revelaron nuevos algoritmos de aprendizaje automático como randon forests y AdaBoost. (Wade 2020)

La idea general detrás del impulso es transformar a los aprendizajes débiles en aprendizajes fuertes mejorando iterativamente los errores, la idea clave detrás del aumento de gradiente es usar el descenso de gradiente para minimizar los errores de los residuos, XGBoost es la abreviatura de Extreme Gradient Boosting, la parte extrema se refiere a empujar los límites de la

computación para lograr ganancias en precisión y velocidad la creciente popularidad de XGBoost se debe en gran parte a su éxito sin precedentes en las competiciones de Kaggle. En las competiciones de Kaggle, los competidores construyen modelos de aprendizaje automático en un intento de hacer las mejores predicciones y ganar lucrativos premios en efectivo, en comparación con otros modelos, XGBoost ha estado aplastando a la competencia. Comprender los detalles de XGBoost requiere comprender el panorama del aprendizaje automático en el contexto del aumento de gradiente. Para pintar una imagen completa, comenzamos por el principio, con los conceptos básicos del aprendizaje automático.(Wade 2020).

2.3.4 OSIPTEL

El Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL) es un organismo técnico especializado del Estado Peruano que regula y supervisa el mercado de servicios públicos de telecomunicaciones; y vela por los derechos del usuario fue creado el 8 de noviembre de 1991 y tiene autonomía técnica, económica, financiera, funcional y administrativa. Está adscrito a la Presidencia del Consejo de Ministros (OSIPTEL 2020).

El OSIPTEL también brinda orientación a la ciudadanía sobre sus derechos y obligaciones como usuarios de los servicios públicos de telecomunicaciones, así como también sobre el procedimiento de reclamos, entre otros trámites.(OSIPTEL 2020)

2.3.5 Atención de Calidad.

La atención de calidad es uno de los aspectos tonamos en cuenta por la regulación, también por la competencia entre empresas operadoras debido que les permite captura o retener a clientes debido a que es uno de los factores que los clientes toman en cuenta a la hora de decidirse o cambiar de una operadora móvil. (OSIPTEL 2021).

2.3.6 Escalas Likert

La Escala de Likert es una escala de calificación, el número de escalas diferentes utilizadas en cualquier estudio varía de uno a siete. Casi todas las escalas son escalas de tipo Likert. La mayoría utiliza 4 o 5 puntos de escala. La mayoría de ellos comienzan con 0 o 1, el extremo negativo de la escala, y progresan hasta 5 o 7, el extremo positivo. (Hartley 2014)

2.3.7 Recursive Feature Elimination RFE

A la hora de procesar datos es muy importante la selección de variables debido a que influyen directamente en el tiempo y la capacidad computación por lo cual el método Recursive Feature Elimination o RFE es una técnica que permite la reducción de estas variables a fin de optimizar mejor los recursos. (Masso and Granitto 2014)

2.3.8 Shrinkage

Según (Rodrigo 2016) este algoritmo lo que realiza es el ajuste del modelo con todos sus predictores aplicando un método que lleva que tiendan a cero a las estimaciones de los cocientes

2.3.8.1 Lasso (L1): Según (Rodrigo 2016) este método aproxima a cero los coeficientes, llegando a excluir predictores. Ambos métodos están especialmente indicados para situaciones en las que hay un mayor número de predictores que de observaciones

2.3.8.2 Ridge (L2): Según (Rodrigo 2016) este método aproxima a cero los coeficientes de los predictores, pero sin llegar a excluir ninguno.

2.3.8.3 Shrinkage L1+L2 (ElasticNet): los métodos muestran performances de acuerdo al escenario en el cual son usados, pudiendo ser en ocasiones de performance similar; usualmente cuando una pequeña porción de predictores de entre los considerados tienen coeficientes normalizados sustanciales y los demás poseen valores similares o iguales a cero, Lasso produce mejores modelos. Si, por el contrario, los predictores considerados tienen coeficientes valores diferentes a cero y de la misma magnitud, Ridge Regression tiende a ser la mejor opción. (Rodrigo 2016)

2.3.9 Imputación de datos

Uno de los problemas frecuentes con el que se topan los investigadores es la calidad de datos dentro estas esta una base de datos incompleta, problema

que afecta directamente en el rendimiento de los algoritmos. para lo cual desde hace mucho tiempo existen técnicas cuyo objetivo es completar atenuar este problema de datos faltantes; estas son conocidas como técnicas de imputación de datos y durante las últimas décadas se han desarrollado procedimientos como la eliminación de datos (Listwise), el pareo de observaciones (Pairwise), el método de medias y el hot-deck.(Medina and Galván 2007)

2.3.10 Dataset.

Data set (conjunto de datos) es un registro de datos tabulados (filas y columnas) correspondiente a una o más tablas de la base de datos, donde las columnas representan una variable particular y las filas a un registro o etiqueta, este conjunto de datos enumera los valores de cada una de las variables. (Gonzales 2013).

2.3.11 Overfitting.

El Overfitting (sobreajuste) es un concepto en ciencia de datos, acontece cuando un algoritmo estadístico encaja perfectamente con sus valores de entrenamiento, cuando esto sucede el modelo desafortunadamente no trabaja con precisión contra datos invisibles, negando su propósito; la generalización de un algoritmo a nuevos datos es, en última instancia, lo que nos permite utilizar modelos de Machine Learning todos los días para clasificar o predecir datos. Cuando se construyen algoritmos de aprendizaje automático, aprovechan un conjunto de datos de muestra para entrenar el modelo. Sin embargo, cuando el modelo se entrena durante demasiado tiempo en datos de muestra o cuando el modelo es demasiado complejo, puede comenzar a aprender el "ruido", o

información irrelevante, dentro del conjunto de datos. Cuando el modelo memoriza el ruido y se ajusta demasiado al conjunto de entrenamiento, el modelo se "sobre ajusta" y no puede generalizar bien a los nuevos datos. Si un algoritmo fracasa en generalizar bien a nuevos datos, no podrá realizar las tareas de clasificación o predicción para las que estaba destinado. Las bajas tasas de error y una alta varianza son buenos indicadores de sobreajuste. Para evitar este tipo de comportamiento, parte del conjunto de datos de entrenamiento generalmente se reserva como el "conjunto de pruebas" para verificar si hay sobreajuste. Si los datos de entrenamiento tienen una tasa de error baja y los datos de prueba tienen una tasa de error alta, indica un sobreajuste.(Baştanlar and Özuysal 2014).

2.3.12 Google colab

(Google 2021) Colaboratory, también llamado "Colab", te permite ejecutar y programar en Python en tu navegador ya seas estudiante, científico de datos o investigador de IA con las siguientes ventajas:

No requiere configuración

Da acceso gratuito a GPUs

Permite compartir contenido fácilmente

2.3.13 Operadores Móviles de Perú.

El Perú tiene 4 operadores móviles físicos con más de 40 millones de líneas móviles activas cuya participación en el mercado a diciembre de 2020 es

como sigue Telefónica 30,74%, seguida por Claro 28,38%, Entel 22,53%, Viettel 17,98% y las OMVs 0,38%.(OSIPTEL 2020)

2.3.14 Calidad de Servicio.

Según OSIPTEL y la Unión Internacional de Telecomunicaciones (UIT-T), conforme a la norma E.800; hablar de calidad involucra desde la parte técnica hasta la gestión y eficiencia que brinda cada proveedor de servicios en telecomunicaciones.(OSIPTEL 2021)

III METODOLOGIA

3.1 Hipótesis general:

Es posible desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

3.2 Hipótesis específicas:

1) Es posible desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación.

2) Se podrá desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

3.3 Identificación de variables:

3.3.1 Variable dependiente.

Calidad de servicio.

3.3.2 Variables independientes.

Indicadores de atención.

- Año
- Mes
- Número de horas sin sistema de atención al mes
- Número total de horas de atención al mes
- Tasa de caída del sistema de atención CSA%
- Número de usuarios que desistieron de la atención al mes
- Número total de usuarios atendidos al mes
- Deserción en atención presencial DAP (%)
- Número de llamadas no finalizadas por el usuario
- Número total de llamadas atendidas
- Corte de la atención telefónica CAT (%)
- Reclamos
- Bajas
- Consultas.
- Altas.

3.4 Nivel de investigación.

La presente investigación es correlacional, descriptiva y explicativa.

Es correlacional por que mide el grado de relación que existe entre las variables metodológicas planteadas.

Es descriptiva, porque no se manipula ninguna de las variables metodológicas estas solo se observan y se describen.

Es explicativa, por que manipula las variables independientes para ver el efecto que provoca en la variable dependiente.

3.5 Tipo de investigación.

Aplicado: porque se usa conocimientos y teorías o de investigación básica para resolver un problema existente.

Cuantitativo: porque genera datos o información numérica que puede ser trabajado de manera estadística.

3.6 Unidad de análisis

Repositorio OSIPTEL (OSIPTEL 2020)

3.7 Población de estudio

Los datos con los que se realiza esta investigación son obtenidos del repositorio de OSIPTEL un total 44 meses que van desde setiembre del 2014 a abril del 2018.

3.8 Tamaño de muestra

En esta investigación no se tomará una muestra, puesto que se trabajará con el total de datos obtenidos.

IV RESULTADOS Y DISCUSIÓN

4.1 Análisis, interpretación y discusión de resultados

4.1.1 Recolección de datos

Todo el data set utilizado para la presente investigación fue obtenida del repositorio de OSIPTEL esta data es reportada por las propias operadoras en concordancia con la normativa de requerimientos de información periódica que tiene este organismo, para el análisis se tomaron en cuenta reportes de 44 meses que van desde setiembre del 2014 a abril del 2018 cuyos indicadores de atención con los siguientes:

- Año
- Mes
- Número de horas sin sistema de atención al mes
- Número total de horas de atención al mes
- Tasa de caída del sistema de atención CSA%
- Número de usuarios que desistieron de la atención al mes
- Número Total de usuarios atendidos al mes
- Deserción en atención presencial DAP (%)
- Número de llamadas no finalizadas por el usuario
- Número total de llamadas atendidas
- Corte de la atención telefónica CAT (%)
- Reclamos

- Bajas
- Consultas
- Altas.

Toda esta información esta agrupada por operador móvil, que para el presente estudio son: MOVISTAR, CLARO y ENTEL adicionalmente poder utilizar este método de evaluación de un modelo de clasificación separamos nuestra data de entrenamiento de cada uno de los 3 operadores en 2 grupos (TRAIN 80% y TEST 20%).

Todo en procesamiento se realizó en Google Colab cuyas características encontraran anexas a este documento, de la misma forma también encontraran el código que fue utilizado para realizar el entrenamiento de aprendizaje automático en lenguaje Python.

4.1.1.1 Movistar

Tabla 1 *Data set Movistar.*

Año[a]	Mes[b]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA %[c]	Nº de usuarios que desistieron de la atención al mes	Nº Total de usuarios atendidos al mes	DAP %[d]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %	Reclamos	Bajas	Consultas	Altas
1	9	0.00	8,529	0	4,370	99,252	0.04402934	88,128	1,677,328	0.0525	0.76	0.79	0.78	0.95
1	10	0.00	26,210	0	26,101	596,323	0.0437699	326,554	5,340,026	0.0612	0.74	0.76	0.78	0.94
1	11	0.00	25,691	0	29,482	585,633	0.05034211	249,708	4,016,975	0.0622	0.72	0.7	0.74	0.92
1	12	10.56	26,503	0.0004	32,598	607,136	0.05369143	308,521	4,263,226	0.0724	0.69	0.67	0.71	0.93
1	1	0.00	26,881	0	29,384	613,609	0.04788717	347,867	4,100,981	0.0848	0.67	0.69	0.74	0.91
1	2	244.16	25,220	0.0097	28,479	602,019	0.04730582	285,814	3,781,024	0.0756	0.72	0.71	0.75	0.91
1	3	0.00	27,370	0	29,126	684,990	0.04252033	342,469	3,568,126	0.096	0.72	0.75	0.78	0.9
1	4	0.00	27,014	0	27,203	602,531	0.04514788	214,585	4,059,290	0.0529	0.67	0.64	0.76	0.9
1	5	0.00	27,349	0	25,829	600,126	0.0430393	170,701	4,006,541	0.0426	0.77	0.69	0.79	0.93
1	6	0	27,166	0	18,368	595,780	0.03083017	200,159	4,059,955	0.0493	0.78	0.69	0.81	0.94
1	7	0.00	26,685	0	19,793	657,470	0.0301048	233,864	4,153,131	0.0563	0.77	0.66	0.79	0.93

1	8	0.00	23,709	0	18,708	646,523	0.02893633	254,517	4,405,683	0.0578	0.76	0.66	0.81	0.93
2	9	0.00	27,070	0	15,582	620,359	0.02511771	219,632	4,103,962	0.05351706	0.81	0.8	0.86	0.92
2	10	0.00	27,547	0	14,912	619,995	0.02405181	219,863	4,202,665	0.05231514	0.83	0.84	0.85	0.92
2	11	0.00	26,219	0	13,119	600,068	0.02186252	206,168	4,074,978	0.05059365	0.82	0.83	0.86	0.95
2	12	0.00	26,963	0	15,853	621,821	0.02549448	227,955	4,049,966	0.05628566	0.74	0.76	0.78	0.94
2	1	61.43	25,966	0.0024	13,763	604,852	0.02275433	207,384	3,996,473	0.05189176	0.79	0.84	0.83	0.95
2	2	0.00	25,839	0	13,816	614,348	0.02248888	274,939	4,006,144	0.06862934	0.77	0.79	0.82	0.93
2	3	0.00	27,896	0	13,672	596,854	0.02290677	317,794	4,204,165	0.07559028	0.8	0.79	0.82	0.9
2	4	0.00	27,896	0	17,721	586,349	0.03022261	171,685	3,954,778	0.04341204	0.81	0.79	0.84	0.94
2	5	0	27,007	0	14,862	589,444	0.02521359	220,583	4,159,508	0.05303103	0.84	0.79	0.83	0.94
2	6	0.00	26,911	0	9,592	545,629	0.01757971	172,762	3,799,637	0.04546803	0.78	0.69	0.8	0.88
2	7	0.00	27,886	0	9,897	546,399	0.01811314	184,055	3,734,124	0.04929001	0.75	0.66	0.78	0.85
2	8	107.28	29,277	0.0037	10,315	612,868	0.0168307	219,841	3,807,484	0.05773918	0.56	0.52	0.71	0.9
3	9	0.00	29,181	0	20,873	690,478	0.03022978	173,470	3,644,319	0.0476	0.65	0.59	0.72	0.86
3	10	0.00	28,061	0	16,073	610,492	0.02632795	155,158	3,308,653	0.0469	0.82	0.7	0.83	0.9
3	11	0.00	27,115	0	16,972	610,781	0.02778737	101,186	2,868,436	0.0353	0.83	0.47	0.82	0.88
3	12	0.00	27,450	0	18,254	663,778	0.02750016	154,084	3,530,983	0.0436	0.84	0.93	0.81	0.87
3	1	155.00	27,144	0.0057	22,692	682,606	0.03324319	164,422	3,783,442	0.0435	0.69	0.78	0.74	0.85

3	2	0.00	25,150	0	19,545	590,145	0.03311898	211,333	3,385,623	0.0624	0.73	0.82	0.81	0.91
3	3	0.00	28,189	0	17,624	621,703	0.02834794	174,786	3,531,808	0.0495	0.81	0.65	0.86	0.94
3	4	0.00	22,252	0	17,624	621,703	0.02834794	104,329	3,384,516	0.0308	0.7	0.48	0.85	0.92
3	5	0.00	27,997	0	16,966	640,729	0.02647921	94,350	3,475,356	0.0271	0.81	0.83	0.85	0.92
3	6	0.00	27,279	0	12,288	578,805	0.02122995	78,796	3,330,663	0.0237	0.87	0.92	0.9	0.95
3	7	0.00	27,327	0	10,572	588,653	0.01795965	60,964	3,407,729	0.0179	0.86	0.85	0.86	0.93
3	8	0.00	27,323	0	9,889	593,356	0.01666622	52,744	3,502,438	0.0151	0.8	0.72	0.87	0.93
4	9	0.00	27,275	0	8,283	562,080	0.01	22,807	3,328,539	0.0069	0.86	0.94	0.91	0.95
4	10	0.00	28,105.5	0	7,455	550,314	0.01	8,856	3,151,060	0.0028	0.88	0.94	0.92	0.96
4	11	0.00	28,063.5	0	9,220	528,507	0.02	9,064	3,136,896	0.0029	0.79	0.89	0.82	0.9
4	12	0.00	27,990.5	0	10,219	557,849	0.02	9,881	2,946,002	0.0034	0.8	0.65	0.84	0.88
4	1	0.00	27,958.0	0	9,180	577,055	0.02	10,648	3,041,823	0.0035	0.8	0.82	0.87	0.92
4	2	0.00	25,390.0	0	8,349	557,021	0.01	10,789	2,778,951	0.0039	0.8	0.87	0.87	0.92
4	3	0.00	27,967.0	0	11,072	612,602	0.02	11,287	3,286,926	0.0034	0.71	0.77	0.79	0.86
4	4	0.00	26,635.5	0	13,422	629,857	0.02	11,472	3,057,883	0.0038	0.7	0.82	0.84	0.88

Fuente: Normativa de Requerimientos de Información Periódica (OSIPTEL 2021)

Nota:

[a] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[b] **Mes:** Correspondiente a un periodo mensual.

[c] **CSA%:** Tasa de caída del sistema de atención, es la relación que existe entre el total de horas sin sistema para la atención y el total de horas en que el sistema debió estar operativo

[d] **DAP (%):** Deserción en atención presencial, es la relación que existe entre el número que usuarios que llego a la oficina saco su ticket y se macho antes de ser atendido.

[e] **CAT (%):** " Corte de la atención telefónica, indica el porcentaje de llamadas no finalizadas por el usuario.

4.1.1.2 Claro

Tabla 2 *Data set Claro.*

Año[a]	Mes[b]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA %[c]	Nº de usuarios que desistieron de la atención al mes	Nº Total de usuarios atendidos al mes	DAP %[d]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %[e]	Reclamos	Bajas	Consultas	Altas
1	9	15.12	18,518	0.082%	41,584	493,360	8.43%	118,385	3,124,089	3.79%	64%	62%	66%	75%
1	10	26.37	18,592	0.142%	37,908	506,790	7.48%	123,382	3,391,270	3.64%	73%	66%	66%	64%
1	11	3.58	18,637	0.019%	37,922	508,222	7.46%	117,383	3,246,056	3.62%	71%	65%	63%	63%
1	12	0.62	21,050	0.003%	39,301	518,982	7.57%	11,756	3,382,625	3.48%	74%	68%	65%	67%
1	1	0.83	19,744	0.004%	35,969	506,900	7.10%	100,277	3,505,628	2.86%	76%	72%	72%	70%
1	2	8.25	18,531	0.045%	26,550	475,277	5.59%	61,818	3,461,584	1.79%	80%	76%	77%	76%
1	3	1.52	20,931	0.007%	25,004	488,132	5.12%	74,489	3,694,472	2.02%	79%	75%	75%	76%
1	4	3.70	19,992	0.019%	17,778	446,492	3.98%	56,881	3,471,795	1.64%	81%	78%	80%	79%

1	5	0	20,780	0.000%	18,789	434,213	4.33%	49,189	3,527,899	1.39%	79%	77%	76%	78%
1	6													
1	7	7.18	20,584	0.035%	16,714	441,291	3.79%	28,800	3,239,626	0.89%	81%	79%	80%	78%
1	8	0.68	20,699	0.003%	16,631	409,005	4.07%	26,790	3,368,573	0.80%	79%	78%	78%	76%
2	9	13.07	20,791	0.063%	16,614	447,039	3.72%	28,200	3,274,924	0.86%	84%	83%	84%	85%
2	10	2.65	20,590	0.013%	18,120	444,409	4.08%	30,046	3,394,403	0.89%	80%	82%	80%	83%
2	11	1.84	20,454	0.009%	20,186	448,584	4.50%	27,794	3,094,064	0.90%	79%	79%	79%	81%
2	12	11.57	21,971	0.053%	24,565	487,204	5.04%	32,331	3,188,555	1.01%	76%	76%	74%	77%
2	1	0.17	20,971	0.001%	21,151	463,831	4.56%	31,975	3,083,228	1.04%	79%	76%	77%	78%
2	2	0.00	20,395	0.000%	25,670	461,846	5.56%	28,301	2,914,392	0.97%	71%	71%	70%	73%
2	3	4.80	21,422	0.022%	25,090	475,539	5.28%	31,074	2,962,227	1.05%	73%	74%	71%	75%
2	4	1.10	21,399	0.005%	21,751	475,211	4.58%	40,603	2,851,686	1.42%	76%	76%	73%	76%

2	5	13.37	22,022	0.061%	20,427	520,146	3.93%	27,667	2,909,221	0.95%	79%	77%	78%	79%
2	6	1.46	21,349	0.007%	15,697	482,379	3.25%	23,314	2,866,739	0.81%	83%	82%	84%	83%
2	7	3.42	21,863	0.016%	15,550	466,933	3.33%	28,001	2,941,578	0.95%	80%	80%	83%	81%
2	8	82.26	22,485	0.366%	17,163	488,156	3.52%	45,691	3,178,351	1.44%	77%	78%	81%	79%
3	9	16.14	22,370	0.072%	18,669	649,167	3.40%	44,135	3,057,234	1.44%	76%	76%	80%	78%
3	10	7.90	22,196	0.036%	19,539	507,969	3.85%	32,328	2,764,541	1.17%	82%	82%	84%	84%
3	11	13.25	21,569	0.061%	16,667	504,509	3.30%	28,163	2,708,331	1.04%	86%	85%	88%	87%
3	12	29.32	22,254	0.132%	23,895	518,464	4.61%	25,228	2,666,419	0.95%	80%	79%	82%	80%
3	1	13.54	22,409	0.060%	21,331	521,947	4.09%	21,419	2,925,380	0.73%	83%	82%	86%	84%
3	2	13.58	20,480	0.066%	19,271	465,185	4.14%	19,006	2,721,997	0.70%	84%	83%	85%	83%
3	3	15.68	22,765	0.069%	17,657	544,011	3.25%	17,873	2,831,259	0.63%	87%	87%	89%	88%
3	4	11.17	20,895	0.053%	15,561	476,504	3.27%	13,984	2,489,875	0.56%	88%	88%	90%	90%

3	5	12.28	22,374	0.055%	21,938	535,544	4.10%	20,078	2,643,379	0.76%	83%	83%	85%	84%
3	6	2.97	21,577	0.014%	22,934	471,505	4.86%	13,129	2,638,914	0.50%	81%	81%	82%	83%
3	7	3.60	21,869	0.016%	19,952	458,687	4.35%	11,814	2,568,330	0.46%	79%	80%	79%	80%
3	8	3.15	22,655	0.014%	16,681	461,356	3.62%	11,236	2,663,202	0.42%	79%	80%	79%	81%
4	9	1.00	22,135	0.005%	13,066	451,604	2.89%	11,106	2,576,963	0.43%	82.70%	83.98%	82.05%	84.47%
4	10	9.53	22,462.25	0.042%	12,246	445,719	2.75%	11,093	2,648,754	0.42%	83.61%	83.32%	84.03%	86.27%
4	11	5.67	22,134	0.026%	12,479	438,504	2.85%	11,550	2,579,743	0.45%	85.06%	85.42%	85.76%	86.63%
4	12	5.28	22,413.5	0.024%	19,828	444,416	4.46%	12,860	2,597,373	0.50%	77.52%	78.39%	75.82%	76.60%
4	1	1.20	22,276	0.005%	20,021	428,153	4.68%	12,344	2,733,556	0.45%	78.81%	79.63%	78.27%	80.15%
4	2	1.38	20,137	0.007%	17,928	384,087	4.67%	11,576	2,544,368	0.45%	79.65%	80.36%	77.92%	82.46%
4	3	2.07	21,792.5	0.009%	17,781	411,650	4.32%	11,564	2,551,255	0.45%	78.51%	78.38%	80.18%	83.36%
4	4	1.57	21,323	0.01%	15,484	415,825	3.72%	11,036	1,459,441	0.45%	80.92%	83.33%	83.66%	86.02%

Fuente: Normativa de Requerimientos de Información Periódica (OSIPTEL 2021)

Nota:

[a] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[b] **Mes:** Correspondiente a un periodo mensual.

[c] **CSA%:** Tasa de caída del sistema de atención, es la relación que existe entre el total de horas sin sistema para la atención y el total de horas en que el sistema debió estar operativo

[d] **DAP (%):** Deserción en atención presencial, es la relación que existe entre el número que usuarios que llego a la oficina saco su ticket y se macho antes de ser atendido.

[e] **CAT (%):** " Corte de la atención telefónica, indica el porcentaje de llamadas no finalizadas por el usuario.

4.1.1.3 Entel

Tabla 3 *Data set Entel.*

Año[a]	Mes[b]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA%[c]	Nº de usuarios que desistieron de la atención al mes	Nº Total de usuarios atendidos al mes	DAP %[d]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %[e]	Reclamos	Bajas	Consultas	Altas
1	9	6.43	9,429	0.000682292	4,989	51,894	0.096138282	50,182	338,717	0.148153178	0.631698974	0.715072084	0.656095591	0.808228896
1	10	2.17	9,807	0.000220931	16,910	84,763	0.199497422	21,615	644,556	0.033534712	0.41010101	0.468875502	0.448331431	0.527332906
1	11	3.57	9,293	0.000383822	16,863	86,335	0.195320554	30,124	539,400	0.055847238	0.377196004	0.4375	0.438255851	0.623782571
1	12	3.50	9,527	0.000367396	11,868	74,383	0.159552586	21,069	582,133	0.03619276	0.448554913	0.541642568	0.548162432	0.750687135
1	1	16.18	9,681	0.001671746	10,389	72,254	0.143784427	19,867	492,719	0.040321157	0.468232576	0.515501691	0.56366626	0.762663774
1	2	29.72	8,998	0.003302586	8,932	67,143	0.133029504	15,546	607,591	0.025586291	0.459779502	0.464622642	0.568765172	0.791792802
1	3	1.23	9,868	0.000124983	9,517	73,596	0.129314093	15,617	673,615	0.023183866	0.479005168	0.4788185	0.553953302	0.792964824

1	4	0.00	9,767	0	9,495	73,652	0.128917069	14,123	652,977	0.021628633	0.465691057	0.517429194	0.505488598	0.81965983
1	5	0.00	9,485	0	8,980	78,846	0.113892905	14,074	643,427	0.021873499	0.502797203	0.530924253	0.545835949	0.814851839
1	6	5.65	9,474	0.0005964	7,686	81,675	0.094104683	13,363	600,422	0.022256013	0.534980989	0.510960334	0.573241132	0.828890869
1	7	13.55	9,474	0.001430306	7,357	90,069	0.081681822	17,830	627,107	0.02843215	0.615803815	0.620985979	0.631460233	0.82156952
1	8	13.20	11,080	0.00119139	8,739	110,115	0.079362485	11,143	705,167	0.015801931	0.563169698	0.557408109	0.579453282	0.802262595
2	9	3.65	11,195	0.000397664	9,439	106,450	0.088670737	9,366	728,733	0.012852444	0.495329528	0.496500778	0.563043838	0.802888858
2	10	25.57	11,152	0.002411546	10,538	112,293	0.093843784	8,691	833,348	0.010429016	0.54319262	0.539574853	0.580417121	0.791845226
2	11	26.20	10,884	0.003161734	10,733	123,951	0.086590669	8,254	824,329	0.010012992	0.529267629	0.519534185	0.573255413	0.788873194
2	12	3.65	11,195	0.000397664	9,318	105,572	0.088262039	9,366	728,733	0.012852444	0.495329528	0.496500778	0.563043838	0.802888858
2	1	8.55	10,926	0.000981703	10,255	112,705	0.090989752	15,096	753,705	0.020029056	0.555169418	0.577301476	0.604215485	0.78487368
2	2	2.45	11,049	0.000273776	8,519	116,685	0.073008527	21,807	711,570	0.030646317	0.554206663	0.587490962	0.601625075	0.783870564
2	3	2.92	11,934	0.000309034	7,650	123,288	0.062049835	20,742	743,322	0.027904461	0.627927147	0.613716295	0.657770542	0.865574541

2	4	7.20	11,328	0.000802544	10,098	137,363	0.073513246	21,980	724,338	0.030344949	0.589407314	0.568992655	0.622653149	0.820420235
2	5	8.45	11,651	0.00083636	9,935	144,857	0.06858488	31,627	749,060	0.042222252	0.601727203	0.541233541	0.620197401	0.83899562
2	6	18.65	11,209	0.001769289	8,566	136,845	0.062596368	26,225	717,125	0.036569636	0.582835821	0.551136364	0.655490989	0.855584888
2	7	0.28	10,923	3.24328E-05	9,226	137,435	0.067129916	27,721	750,059	0.036958426	0.568421053	0.522046285	0.595398429	0.836157961
2	8	7.77	11,651	0.000803937	11,983	164,667	0.072771108	31,008	863,904	0.035892877	0.547165633	0.476124704	0.581253328	0.812658635
3	9	6.32	11,651	0.000542157	5,198	175,787	0.029569877	26,224	829,122	0.031628638	0.634547069	0.553491101	0.669473	0.854300761
3	10	1.74	11,651	0.000149343	3,104	155,053	0.020018961	22,771	787,734	0.028906966	0.716680362	0.666408068	0.746599088	0.870173494
3	11	8.32	11,495	0.000723503	2,717	157,561	0.017244115	23,676	801,422	0.029542488	0.787015044	0.783847669	0.815350651	0.882431967
3	12	0.00	11,783	0	3,897	173,154	0.022505977	25,426	855,474	0.029721534	0.720719207	0.722325987	0.746338803	0.843994665
3	1	3.03	11,647	0.000260153	3,601	181,251	0.019867477	27,354	846,245	0.032323972	0.733620072	0.72097144	0.760398066	0.828830874
3	2	2.87	11,938	0.00024014	3,162	178,552	0.017709127	25,198	779,991	0.032305501	0.83235203	0.732129514	0.790221	0.735316322
3	3	33.70	11,938	0.002823037	3,254	182,871	0.017793964	32,364	841,070	0.038479556	0.758600498	0.745641271	0.81010567	0.859790903

3	4	2.67	11,938	0.000223386	3,222	168,281	0.019146547	38,391	785,477	0.048876033	0.773094023	0.758017493	0.796853789	0.845037783
3	5	3.68	11,938	0.000308551	3,857	197,825	0.01949703	31,179	889,714	0.035043846	0.761681635	0.734771432	0.800524789	0.854405498
3	6	1.58	11,938	0.000132635	3,615	180,645	0.020011625	28,243	831,701	0.033958117	0.752563226	0.711588339	0.81993104	0.899706086
3	7	0.00	11,938	0	5,232	190,318	0.027490831	29,615	881,653	0.033590313	0.693251534	0.738069414	0.750595269	0.847227902
3	8	3.02	11,647	0.000259008	5,216	194,235	0.026854069	29,123	920,062	0.031653302	0.767544429	0.826539924	0.791159276	0.83547958
4	9	2.93	11,340.50	0.00025866	3,953	183,241	0.021572683	30,117	833,695	0.036124722	0.820102136	0.812386341	0.849048135	0.884191849
4	10													
4	11													
4	12	0.00	10,911.50	0	5,593	188,532	0.029666051	30,385	973,143	0.031223571	0.74947087	0.74524248	0.769406531	0.864367572
4	1	0.00	11,647	0	4,117	190,757	0.021582432	25,891	867,100	0.029859301	0.794815771	0.790859077	0.814337999	0.886269872
4	2	1.18	7,796	0.000151787	4,023	172,367	0.023339734	24,915	779,504	0.031962633	0.77590659	0.768499856	0.804085461	0.880270229
4	3	0.00	8,143.50	0	3,487	178,672	0.019516208	23,563	808,844	0.029131699	0.811104088	0.813582888	0.824838022	0.866769706

4	4	0.78	7,948	0.000151787	3,564	177,067	0.020127974	15,292	780,906	0.019582383	0.758903561	0.744084137	0.794144722	0.844355617

Fuente: Normativa de Requerimientos de Información Periódica (OSIPTEL 2021)

Nota:

[a] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[b] **Mes:** Correspondiente a un periodo mensual.

[c] **CSA%:** Tasa de caída del sistema de atención, es la relación que existe entre el total de horas sin sistema para la atención y el total de horas en que el sistema debió estar operativo

[d] **DAP (%):** Deserción en atención presencial, es la relación que existe entre el número que usuarios que llego a la oficina saco su ticket y se macho antes de ser atendido.

[e] **CAT (%):** " Corte de la atención telefónica, indica el porcentaje de llamadas no finalizadas por el usuario.

.

4.1.2 Método propuesto

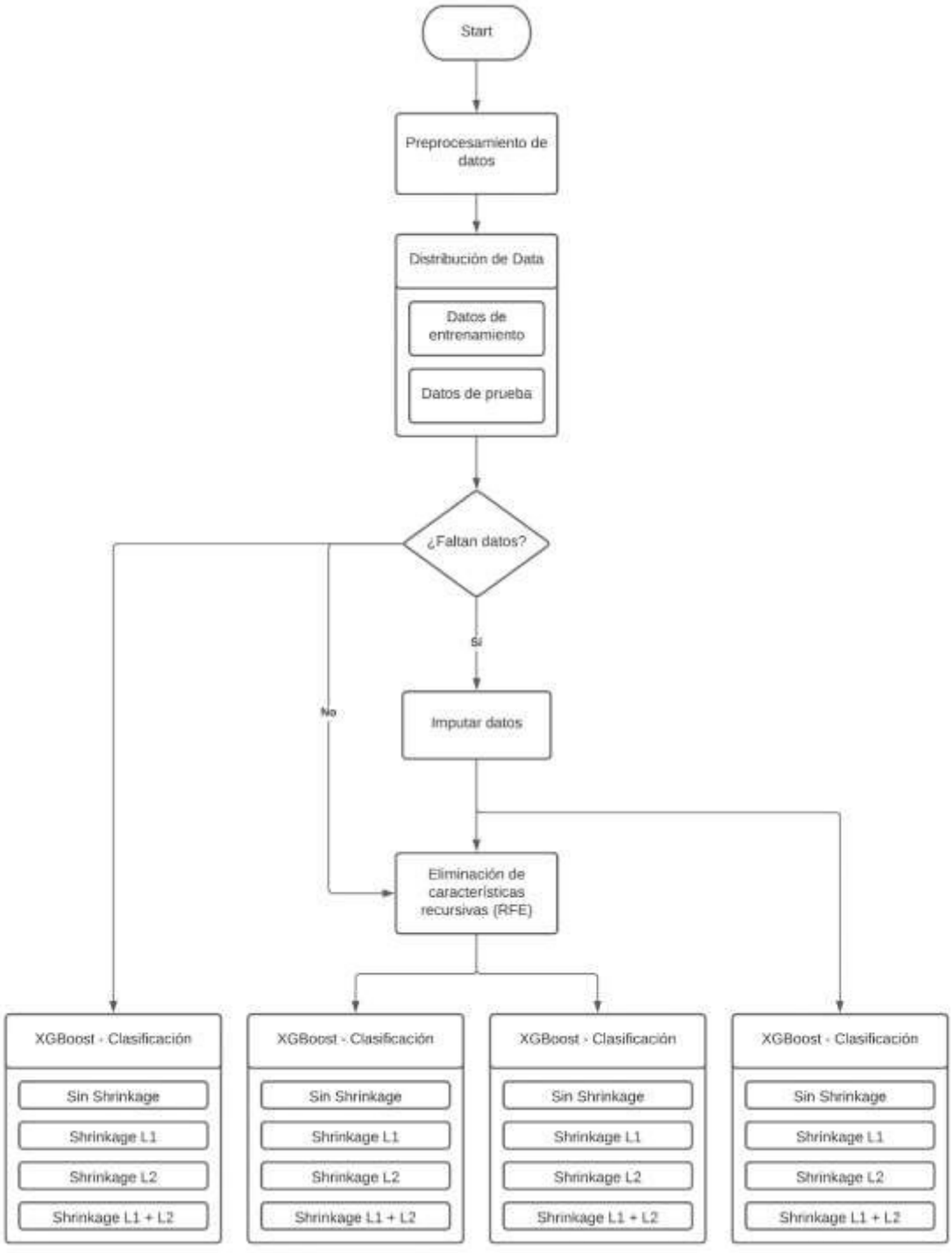

Figura 1 Diagrama de Flujo.

Fuente: Elaboración Propia.

4.1.3 El algoritmo

4.1.3.1 eXtreme Gradient Boosting

XGBoost es un algoritmo que se utiliza en Machine Learning, es decir, con datos tabulares para su predicción o clasificación (de Vito 2017). Su principio de funcionamiento son los árboles de decisión que representan gráficamente las posibles soluciones. Luego pasan por un proceso de ensacado que combina varias predicciones de varios árboles de decisión utilizando un sistema mayoritario. Luego pasan por un proceso de Randon Forest en el que se selecciona aleatoriamente un conjunto de características relevantes para el sistema. Posteriormente pasan por un proceso de Boosting para minimizar los errores de los modelos obtenidos previamente mediante un proceso de impulso de gradiente o de gradiente descendente. (Zolotareva 2021). Por lo tanto, obtiene un algoritmo de aumento de gradiente optimizado a través del procesamiento paralelo, la poda de árboles, el manejo de valores faltantes y la regularización para evitar el sobreajuste y los sesgos.

La función objetivo puede denotarse a través de los dos términos del (Ec 1) donde el primero representa la pérdida de formación.

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (1)$$

y el segundo el término relacionado con la regularización

Los términos y representan los valores actuales y previstos y el término l denota el error entre ellos. El término Ω es la función de regularización para evitar el sobreajuste para los árboles de decisión k. En general, el error se puede determinar a través de:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad (2)$$

El árbol de decisión se puede representar a partir de su complejidad:

$$f_i(x) = \omega_{q(x)}, \omega \epsilon R^T, q: R^d \rightarrow \{1,2, \ldots T\} \quad (3)$$

donde T es el número total de hojas, w es un sub vector de hojas, y la función q asigna cada punto de datos a la hoja correspondiente.

La regularización se puede expresar como:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} \omega_i^2 \quad (4)$$

donde Y y λ son los coeficientes asociados a los términos regulares. El modelo de adición y el algoritmo de pasos hacia adelante proporcionan las funciones de entrenamiento y optimización para el modelo XGBoost (Deng y Lumley 2021) .

El proceso aritmético del modelo se puede expresar en función de los valores predichos como:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \quad (5)$$

Así que finalmente, la función objetivo puede ser reescrita como:

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \quad (6)$$

4.1.3.2 Recursive Feature Elimination RFE

El método Recursive Feature Elimination (RFE) ordena las variables conforme a su importancia que es otorgada por un clasificador; que realiza iteraciones con la finalidad de medir la relevancia de todas las variables desechando la menos importante. Para hacer más eficiente el proceso cada vez que el algoritmo realiza iteraciones va desechando la o las variables menos importantes acelerando el proceso en cada barrido. (Masso and Granitto 2014)

$$r_i = \beta|w_i| + (1-\beta)\frac{R_i}{Q_{S,i}}, \quad (7)$$

4.1.3.3 Data imputation

La imputación de datos es una técnica que básicamente intenta reponer datos perdidos y la imputación de la media, consiste en utilizar la media muestral de los valores accesibles, y están dados por $y_i^* = \bar{y}_r, i \in s_m, Donde$: (Muñoz Rosas and Álvarez Verdejo 2009)

$$\bar{y}_r = \frac{1}{\sum_{i \epsilon s_r} d_i} \sum_{i \epsilon s_r} d_i y_i \quad (8)$$

4.1.3.4 Shrinkage L1+L2 (Elastic Net)

Elastic Net mezcla las regularizaciones L1 y L2. Utilizando el parámetro r y la importancia relativa que matemáticamente está dada por:

$$C = r \cdot Lasso + (1 - r) \cdot Ridge \quad (9)$$

Y para el error cuadrático medio tenemos:

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2 + r \cdot \propto \frac{1}{N}\sum_{j=1}^{N}|w_i| + (1 - r) \cdot \alpha \frac{1}{2N}\sum_{j=1}^{N} w_i^2 \quad (10)$$

4.1.3.5 Shinkage L2 (Ridge)

Así como el conjunto de datos que queremos utilizar para hacer modelos de Machine learning debe seguir la distribución gaussiana definida por su media μ y varianza σ^2 y está representada por $N(\mu, \sigma^2)$, es decir, $X \sim N(\mu,\sigma^2)$ donde X s la matriz de entrada.

Para cualquier punto x_i, la probabilidad de x_i está dada por:

$$P(x_i) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \quad (11)$$

La ocurrencia de cada x_i s independiente de la ocurrencia de otro, la probabilidad conjunta de cada uno está dada por:

$$p(x_1, x_2, ...x_N) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \quad (12)$$

Además, la regresión lineal es la solución que da la máxima probabilidad a la línea de mejor ajuste:

$$P(X \mid \mu) = p(x_1, x_2, ...x_N) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

(13)

Para ello tomamos el logaritmo natural de la función de probabilidad (probabilidad) (L), luego diferenciamos e igualamos cero.

$$ln(P(X \mid \mu)) = ln(p(x_1, x_2, ...x_N)) = \tag{14}$$

$$ln \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$= \sum_{i=1}^{N} \ln\left(\frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}\right) =$$

$$\sum_{i=1}^{N} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \sum_{i=1}^{N} \mid \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} \tag{15}$$

$$\frac{\partial \ln(P(X \mid \mu)}{\partial \mu} = \frac{\partial \sum_{i=1}^{N} \ln\left(\frac{1}{2\pi\sigma^2}\right)}{\partial \mu} - \frac{\partial \sum_{i=1}^{N} \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}{\partial \mu} \tag{16}$$

$$= 0 + \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} \tag{17}$$

$$\frac{\partial \ln(P(X \mid \mu)}{\partial \mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0 \Longrightarrow \mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{18}$$

Tenemos en cuenta aquí es que maximizar la función de probabilidad (probabilidad) L es equivalente a minimizar la función de error E. Además, es gaussiano distribuido con transposición media (w) * X y varianza σ^2

$$y \sim N(\omega^T X, \sigma^2) \tag{19}$$

o

$$y = \omega^T X + \varepsilon \tag{20}$$

Dónde:

$$\varepsilon \sim N(0, \sigma^2)$$

ε es ruido distribuido gaussiano con media cero y varianza σ^2. Los errores son gaussianos y la tendencia es lineal. Para valores nuevos u atípicos, la predicción sería menos precisa para mínimos cuadrados, por lo que utilizaríamos el método de regularización L2. Por lo tanto, modificamos la función de costo y penalizamos los pesos grandes de la siguiente manera.

$$J_{RIDGE} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda |w|^2 \quad (21)$$

Dónde:

$$|w|^2 = w^T w = w_1^2 + w_2^2 + \cdots + w_D^2 \quad (22)$$

Ahora tenemos dos probabilidades:

Posterior:

$$P(Y|X, w) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} exp(-\frac{1}{2\sigma^2}(y_n - w^T x_n)^2) \quad (23)$$

A priori:

$$P(w) = \frac{\lambda}{\sqrt{2\pi}} exp(-\frac{\lambda}{2} w^T w) \quad (24)$$

4.1.3.6 Shinkage L1 (Lasso)

De la misma manera para Lasso

$$J_{LASSO} = \sum_{n=1}^{N} (y_i - \hat{y}_i)^2 + \lambda ||w|| \tag{25}$$

Maximización de la probabilidad

$$P(Y|X,w) = \prod_{n=1}^{N} \frac{1}{2\pi\sigma^2} exp(-\frac{1}{2\sigma^2}(y_n - w^T x_n)^2) \tag{26}$$

y el prior está dado por:

$$P(w) = \frac{\lambda}{2} exp(-\lambda|w|) \tag{27}$$

Para

$$J = (Y - X_w)^T (Y - X_w) + \lambda|w| \tag{28}$$

Y

$$\frac{\partial J}{\partial w} = -2X^T Y + 2X^T Y + 2X^T X w + \lambda sign(w) = 0 \tag{29}$$

Dónde: $sign(\omega) = 1 if x > 0\ y\ -1\ if\ x\ < 0\ and\ 0\ if\ x = 0$

4.2 Pruebas de hipótesis

4.2.1 Matriz de confusión

Una forma de evaluar el modelo, es usando la matriz de confusión, que nos permite una vez terminado el entrenamiento contrastar los resultados ciertos o falsos de la inferencia de nuestro modelo sobre los datos de entrenamiento comparado con la referencia de la data set. esto nos ayudar a medir como se comportará nuestro modelo cuando se lo aplicamos a data nueva. Durante el entrenamiento se aplica una evaluación con un set de datos aleatorias de la data set, para verificar que el algoritmo no tienda al overfitting. La matriz tiene la siguiente estructura:

		Predicción	
		Positive	Negative
Real	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figura 2 Matriz de confusión

Fuente: Elaboración Propia.

Donde:

- TP Positivos reales identificados como positivos por el algoritmo.
- TN Negativos reales identificados como negativos por el algoritmo.
- FN Positivos reales identificados como negativos por el algoritmo
- FP Negativos reales identificados como positivos por el algoritmo

A partir de la matriz de confusión inferimos que data puede ser útil al momento de evaluar el algoritmo.

4.2.1.1 Precisión: Cuantifica el éxito de la predicción considerando los falsos positivos y está dada por:

$$precision = \frac{TP}{TP+FP} \quad (30)$$

4.2.1.2 Exactitud: Qué porcentaje de predicciones fueron correctas y está dado por:

$$exactitud = \frac{TP+TN}{TP+TN+FN+FP} \quad (31)$$

4.2.1.3 Sensibilidad: Qué porcentaje de casos positivos capturó el modelo y este dado por:

$$sensibilidad = \frac{TP}{TP+FN} \quad (32)$$

4.2.1.4 Puntaje F1: Promedio ponderado de precisión y Sensibilidad se calcula de la siguiente manera:

$$puntaje\ F1 = \frac{2*precision*sensibilidad}{precision+sensibilidad} \quad (33)$$

4.2.2 Curva ROC

Característica Operativa del Receptores (ROC) una representación gráfica de la sensibilidad frente a la especificidad (Martínez-Camblor 2007)

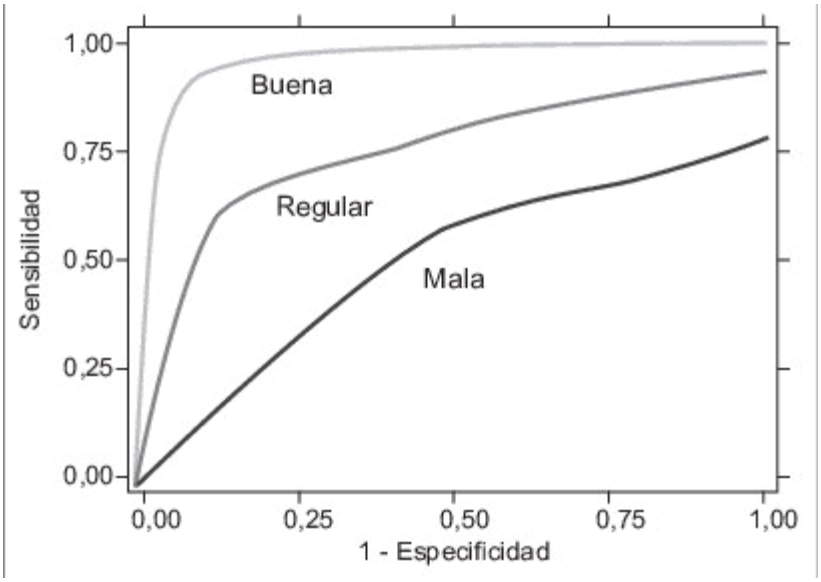

Figura 3 Curva de ROC

Fuente: *(Martínez-Camblor 2007)*

4.3 Presentación de resultados

4.3.1 Movistar

4.3.1.1 Descripción Estadística Data set Movistar.

Tabla 4 Descripción Estadística Data set Movistar.

	AÑO[d]	Mes[e]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA %[f]	Nº de usuarios que desistieron de la atención al mes	Nº total de usuarios atendidos al mes	DAP %[g]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %[h]	Reclamos	Bajas	Consultas	Altas
Count[a]	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44
Mean[b]	2.363636	6.5	13.14614	26560.28	0.000498	16570.84	593611.2	0.027897	165590.3	3668369	0.043076	0.768636	0.748864	0.813636	0.914545
Std[c]	1.080287	3.63126	46.11138	3062.655	0.001774	6985.731	84467.2	0.011106	101745.4	579079.9	0.024298	0.066178	0.112853	0.051539	0.029131
min	1	1	0	8529	0	4370	99252	0.01	8856	1677328	0.0028	0.56	0.47	0.71	0.85
25%	1	3	0	26431.88	0	10507.75	586170	0.02	85795	3330132	0.02625	0.72	0.685	0.78	0.9
50%	2	6.5	0	27155	0	15717.5	602275	0.025911	174128	3757574	0.049295	0.78	0.765	0.82	0.92
75%	3	10	0	27895.5	0	19607	620695	0.031402	222426	4059456	0.05666	0.81	0.8225	0.85	0.94
max	4	12	244.16	29276.5	0.0097	32598	690478	0.053691	347867	5340026	0.096	0.88	0.94	0.92	0.96

Fuente: Elaboración propia.

Nota:

[a] Count: Cantidad de datos.

[b] Mean: Valor medio.

[c] Std: Desviación estándar.

[d] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[e] **Mes:** Correspondiente a un periodo mensual.

[f] **CSA%**: Tasa de caída del sistema de atención.

[g] **DAP %:** Deserción en atención presencial.

[h] **CAT %:** "Corte de la atención telefónica.

4.3.1.2 Resultados con Machine Learning

Tabla 5 Resultados Movistar - XGBoost.

			Precisión	Exactitud	Sensibilidad	Puntaje F1	Curva ROC	M. Confusión
Sin Imputación (mean - media)	Sin RFE	Sin Shrinkage	0.83333333	0.88888889	0.92857143	0.86153846	0.92857143	[[6 1] [0 2]]
		Shrinkage L1	0.67857143	0.77777778	0.67857143	0.67857143	0.67857143	[[6 1] [1 1]]
		Shrinkage L2	0.67857143	0.77777778	0.67857143	0.67857143	0.67857143	[[6 1] [1 1]]
		Shrinkage L1+L2	0.67857143	0.77777778	0.67857143	0.67857143	0.67857143	[[6 1] [1 1]]
	Con RFE*	Sin Shrinkage	0.83333333	0.88888889	0.92857143	0.86153846	0.92857143	[[6 1] [0 2]]
		Shrinkage L1	0.93750000	0.88888889	0.75000000	0.80000000	0.75000000	[[7 0] [1 1]]
		Shrinkage L2	0.83333333	0.88888889	0.92857143	0.86153846	0.92857143	[[6 1] [0 2]]
		Shrinkage L1+L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]

Fuente: Elaboración propia.

* Variables consideradas 7: 'AÑO', 'CSA%', 'DAP%', 'CAT%', 'Bajas', 'Consultas', 'Altas'

4.3.1.3 Diagrama de calor de indicadores de atención Movistar

El algoritmo RFE; es una técnica de preprocesado cuyo objetivo es la selección de variables. Este algoritmo nos permite reducir variables eliminando las que tienen escasa y nula relación con el objetivo del estudio y cuando una o más variables contribuyen con la misma información respecto al objetivo del estudio, en la siguiente figura podemos observar la relación entre las variables.

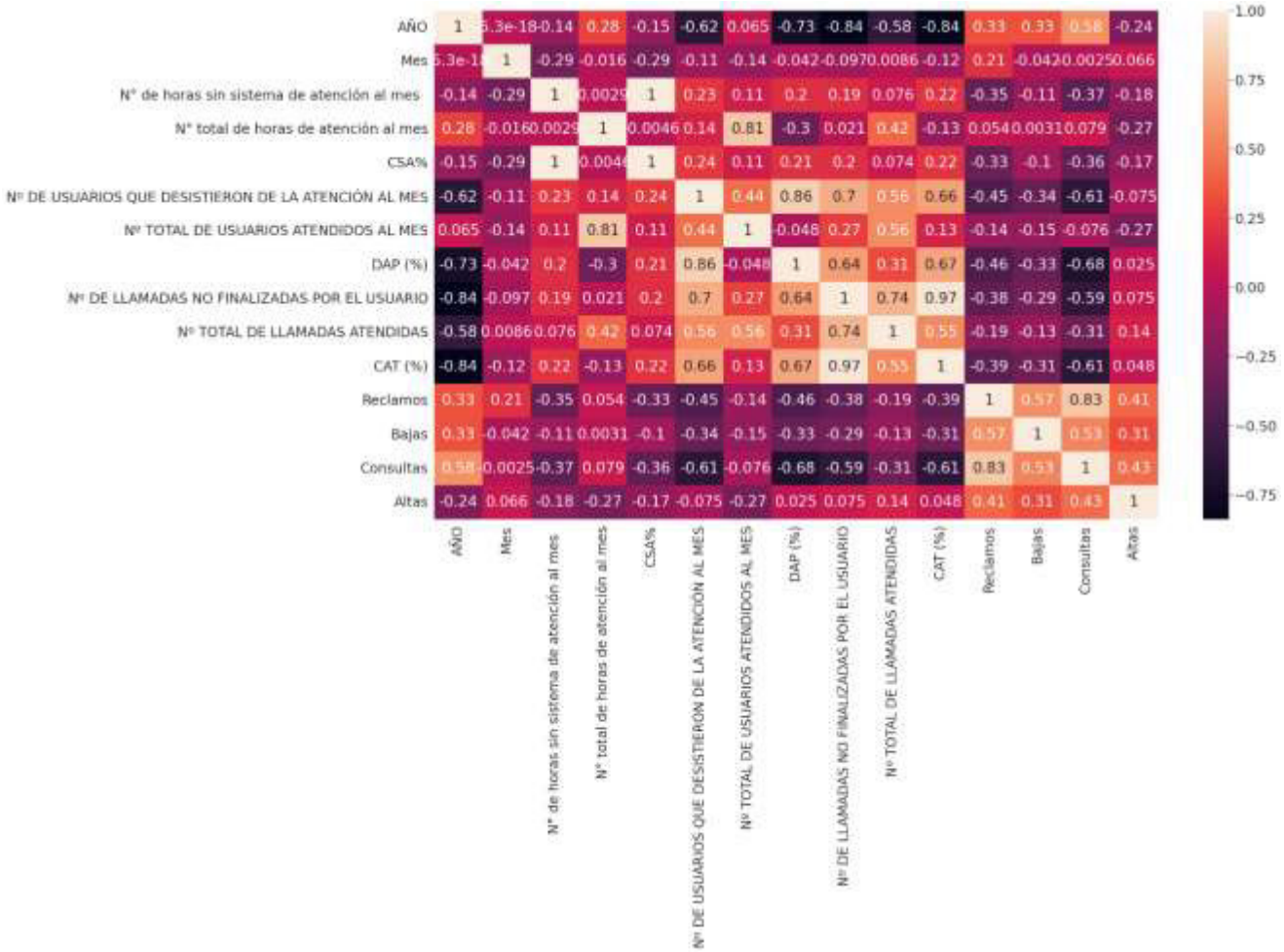

Figura 4 Diagrama de calor de indicadores de atención Movistar

Fuente*:* Elaboración Propia.

4.3.1.4 Resultados comparativos Movistar

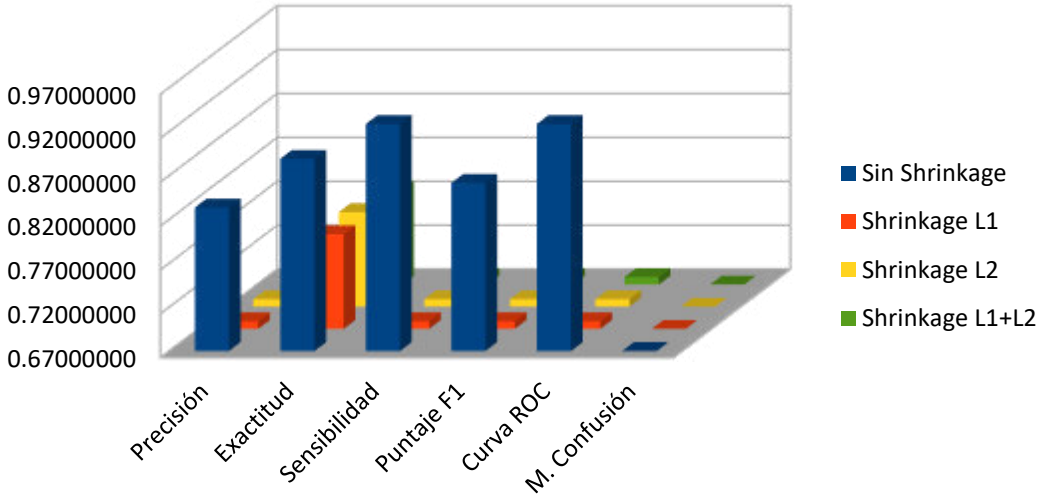

Figura 5 Movistar Sin Imputación - Sin RFE.

Fuente*:* Elaboración Propia.

Como se puede observar en la Figura 5, no se realizó imputación, pues los datos estaban completos, se obtuvo como mejor resultado el modelo Sin Shrinkage con una precisión de 83.33%, una exactitud de 88.89%, una sensibilidad de 92.86 y un puntaje F1 de 86.15%.

Este modelo no utilizo RFE por lo tanto no se eliminó ninguna y se consideraron las 15 propuesta para su evaluación.

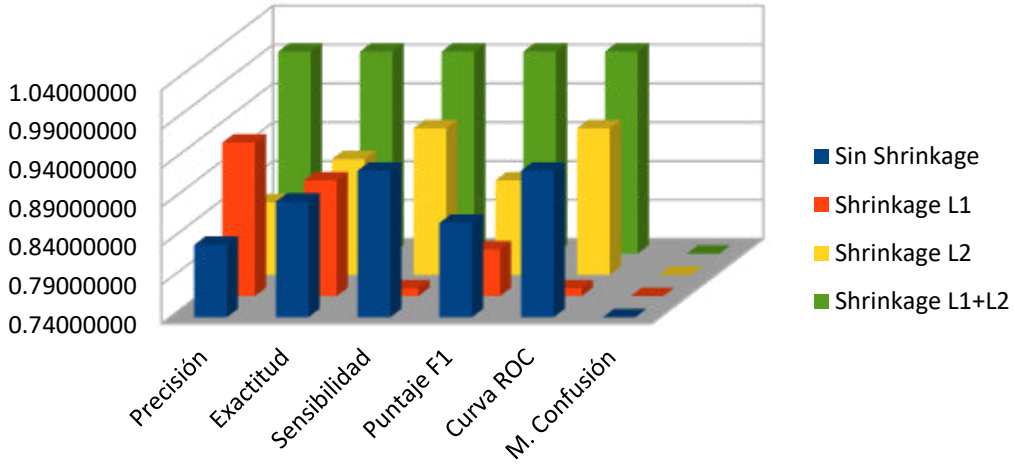

Figura 6 Movistar Sin Imputación - Con RFE

Fuente*:* Elaboración Propia.

Del mismo modo como en la figura 5 en este modelo figura 6 no tiene datos imputados pues los datos estaban completos, se obtuvo como mejor resultado el modelo hibrido RFE + Sin Shrinkage y el modelo RFE + Shrinkage L2 con una precisión de 83.33%, una exactitud de 88.89%, una sensibilidad de 92.86 y un puntaje F1 de 86.15%.

El método propuesto RFE logro reducir el número de variables considerando significantes las siguientes: AÑO, CSA%, DAP %, CAT %, Bajas, Consultas, Altas.

4.3.2 Claro

4.3.2.1 Descripción Estadística Dataset Claro.

Tabla 6 Resumen Estadístico Dataset Claro.

	AÑO[d]	Mes[e]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA %[f]	Nº de usuarios que desistieron de la atención al mes	Nº total de usuarios atendidos al mes	DAP %[g]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %[h]	Reclamos	Bajas	Consultas	Altas
Count[a]	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44
Mean[b]	2.363636	6.5	13.14614	26560.28	0.000498	16570.84	593611.2	0.027897	165590.3	3668369	0.043076	0.768636	0.748864	0.813636	0.914545
Std[c]	1.080287	3.63126	46.11138	3062.655	0.001774	6985.731	84467.2	0.011106	101745.4	579079.9	0.024298	0.066178	0.112853	0.051539	0.029131
min	1	1	0	8529	0	4370	99252	0.01	8856	1677328	0.0028	0.56	0.47	0.71	0.85
25%	1	3	0	26431.88	0	10507.75	586170	0.02	85795	3330132	0.02625	0.72	0.685	0.78	0.9
50%	2	6.5	0	27155	0	15717.5	602275	0.025911	174128	3757574	0.049295	0.78	0.765	0.82	0.92
75%	3	10	0	27895.5	0	19607	620695	0.031402	222426	4059456	0.05666	0.81	0.8225	0.85	0.94
max	4	12	244.16	29276.5	0.0097	32598	690478	0.053691	347867	5340026	0.096	0.88	0.94	0.92	0.96

Fuente: Elaboración propia.

Nota:

[a] Count: Cantidad de datos.

[b] Mean: Valor medio.

[c] Std: Desviación estándar.

[d] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[e] **Mes:** Correspondiente a un periodo mensual.

[f] **CSA%**: Tasa de caída del sistema de atención.

[g] **DAP %:** Deserción en atención presencial.

[h] **CAT %:** "Corte de la atención telefónica.

4.3.2.2 Resultados con Machine Learning

Tabla 7 Resultados Claro. - XGBoost

			Precisión	Exactitud	Sensibilidad	Puntaje F1	Curva ROC	M. Confusión
Sin Imputación (mean - media)	Sin RFE	Sin Shrinkage	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L1	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L1+L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
	Con RFE*	Sin Shrinkage	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L1	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]
		Shrinkage L1+L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[7 0] [0 2]]

Con Imputación (mean - media)	Sin RFE	Sin Shrinkage	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L1	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L2	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L1+L2	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
	Con RFE**	Sin Shrinkage	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L1	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L2	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]
		Shrinkage L1+L2	0.92857143	0.88888889	0.83333333	0.86153846	0.83333333	[[6 0] [1 2]]

Fuente: Elaboración Propia.

Nota:

* Variables Consideras 12: AÑO, Mes, Número de horas sin sistema de atención al mes, Número total de horas de atención al mes, CSA %, Número de usuarios que desistieron de la atención al mes, Número total de usuarios atendidos al mes, DAP %, CAT %, Bajas, Consultas, Altas.

** Variables Consideras 8: Mes, N° de horas sin sistema de atención al mes, CSA %, DAP %, CAT %, Bajas, Consultas, Altas

4.3.2.3 Diagrama de calor de indicadores de atención Claro.

El algoritmo RFE; es una técnica de preprocesado cuyo objetivo es la selección de variables. Este algoritmo nos permite reducir variables eliminando las que tienen escasa y nula relación con el objetivo del estudio y cuando una o más variables contribuyen con la misma información respecto al objetivo del estudio, en la siguiente figura podemos observar la relación entre las variables.

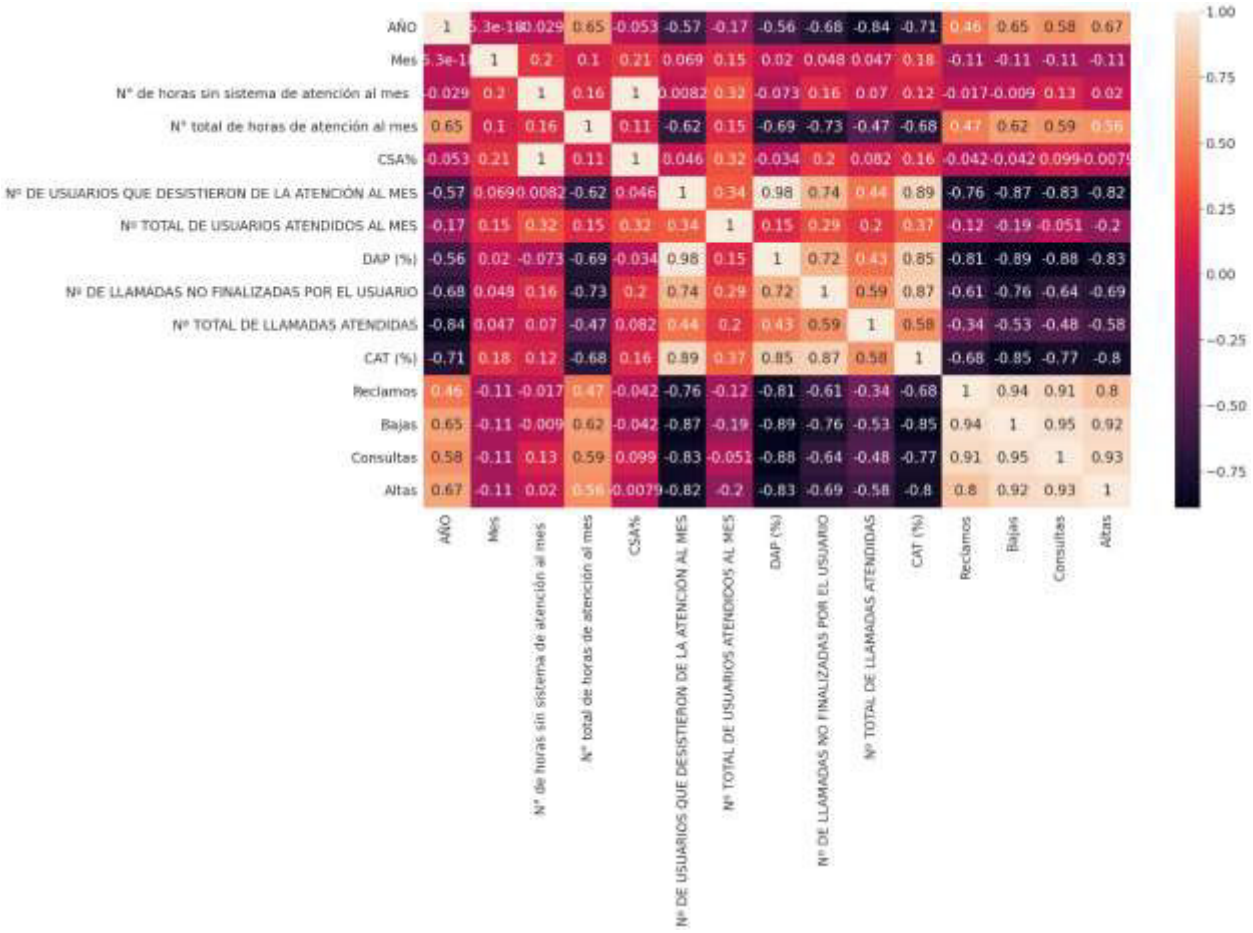

Figura 7 Diagrama de calor de indicadores de atención Claro.

Fuente*:* Elaboración Propia.

4.3.2.4 Resultados comparativos Claro

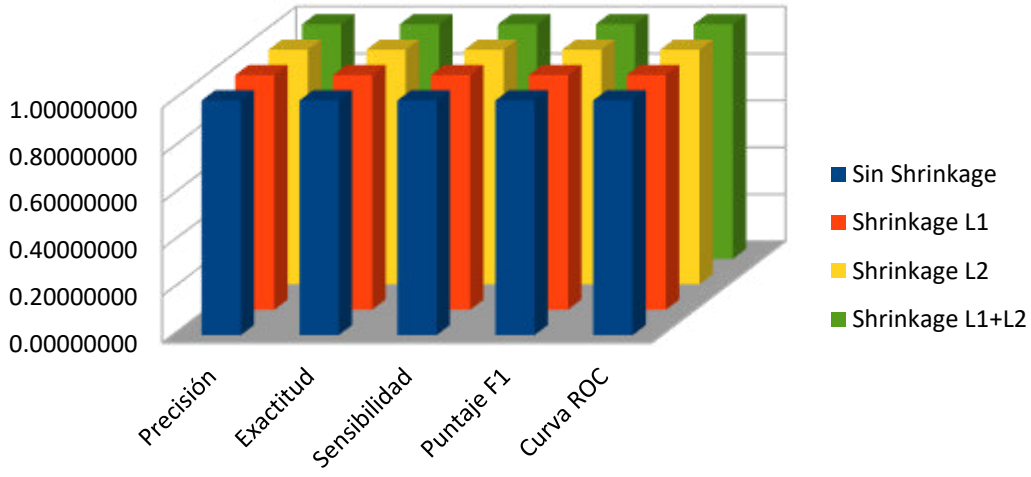

Figura 8 CLARO Sin Imputación - Sin RFE

Fuente*:* Elaboración Propia.

En la figura 8 se muestran los resultados para el modelo sin imputación y sin RFE, en esta notamos que todos los modelos tienen Overfitting que es un concepto en Machine Learning que se da cuando el modelo propuesto coincide con los resultados de su entrenamiento. Cuando esto sucede, el modelo no se puede utilizar ya que habría memorizado.

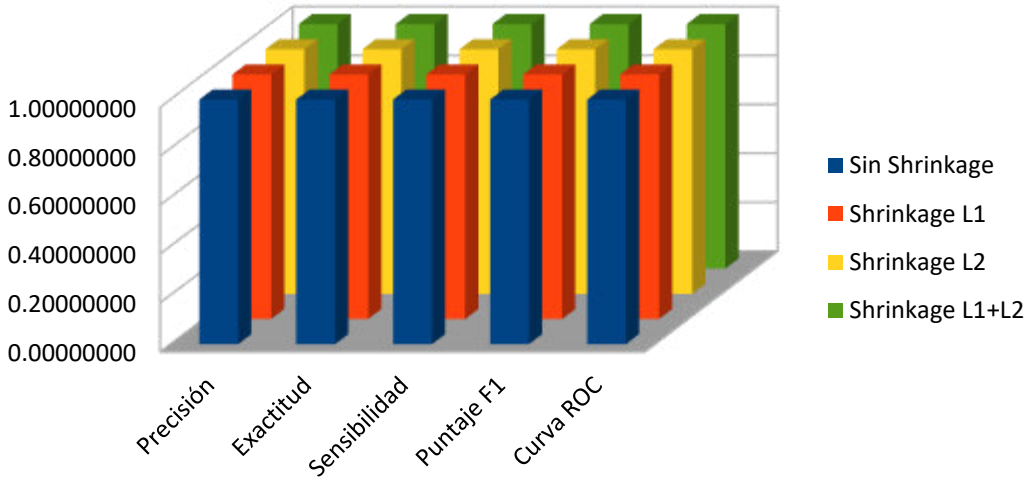

Figura 9 CLARO Sin Imputación - Con RFE.

Fuente: Elaboración Propia.

En la figura 9 podemos observar los resultados para el modelo sin imputación y con RFE, en esta notamos que todos los modelos tienen Overfitting que es un concepto en Machine Learning que se da cuando el modelo propuesto coincide con los resultados de su entrenamiento. Cuando esto sucede, el modelo no se puede utilizar ya que habría memorizado.

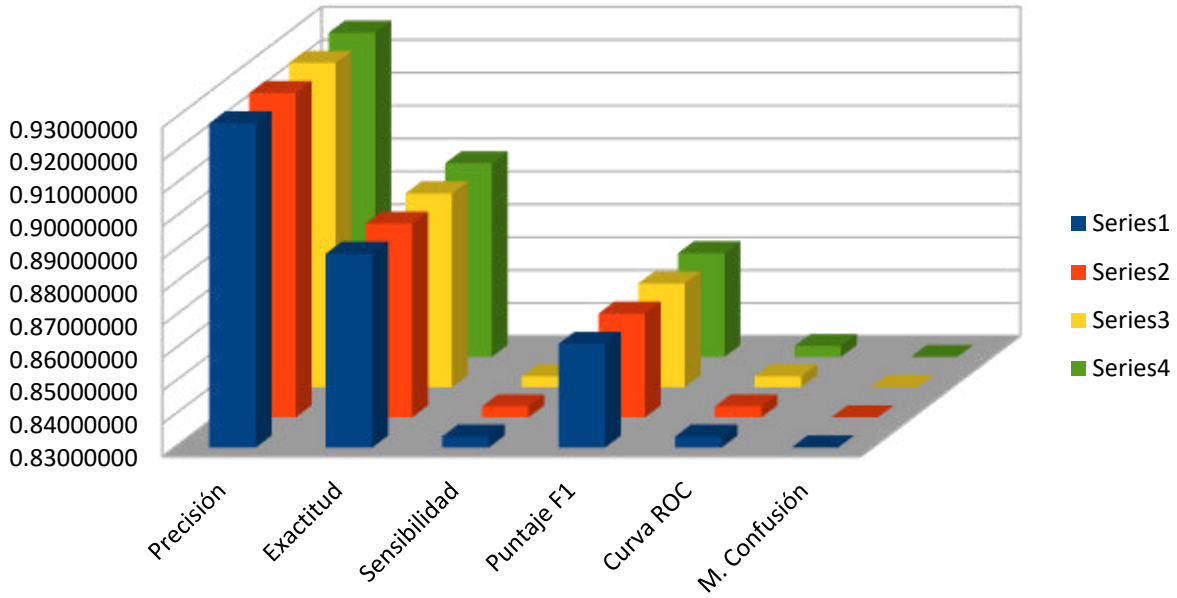

Figura 10 CLARO Con Imputación - Sin RFE

Fuente*:* Elaboración Propia.

En la figura 10, se realizó la técnica de imputación pues los datos no estaban completos no se utilizó el método RFE y los resultados obtenidos muestran que los cuatro algoritmos Sin Shrinkage, Shrinkage L1, Shrinkage L2, Shrinkage L1+L2 obtuvieron una precisión de 92.86%, una exactitud de 88.89%, una sensibilidad de 83.33 y un puntaje F1 de 86.15%.

Este modelo trabajó con las 15 variables debido a que no se utilizó ninguna técnica de RFE.

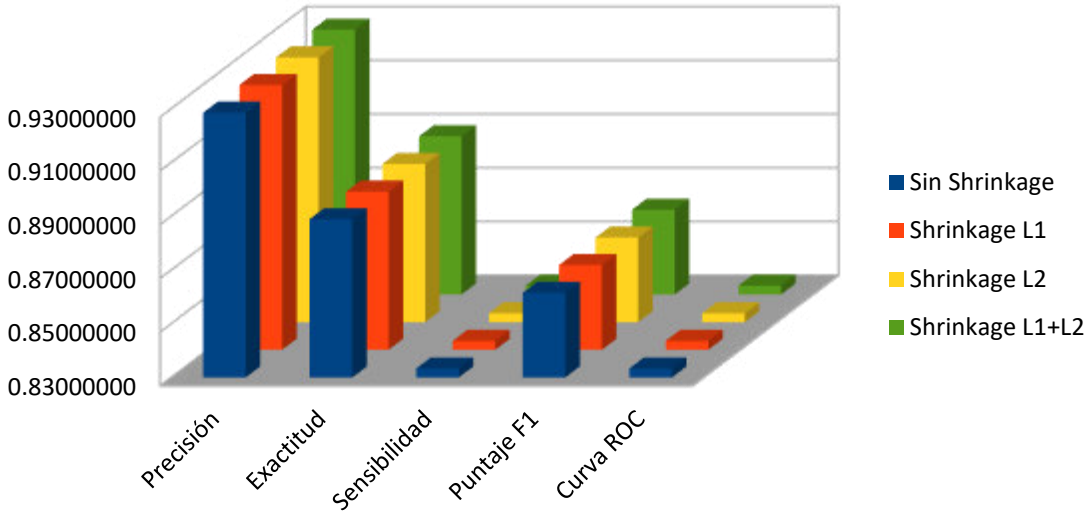

Figura 11 CLARO Con Imputación - Con RFE.

Fuente*:* Elaboración Propia.

En la figura 11, se realizó imputación de datos además se usó el método RFE y los resultados obtenidos muestran que los cuatro algoritmos Sin Shrinkage, Shrinkage L1, Shrinkage L2, Shrinkage L1+L2 obtuvieron una precisión de 92.86%, una exactitud de 88.89%, una sensibilidad de 83.33 y un puntaje F1 de 86.15%.

. También debo mencionar que después de aplicar el método hibrido con RFE se redujo el número de variables tomando en cuenta solo los que tenían significancia real para el modelo manteniéndose las siguientes variables independientes: Mes, N° de horas sin sistema de atención al mes, CSA%, DAP%, CAT%, Bajas, Consultas, Altas

4.3.3 Entel

4.3.3.1 Descripción Estadística Data set Entel.

Tabla 8 Resumen Estadístico Data set Entel.

	AÑO[d]	Mes[e]	N° de horas sin sistema de atención al mes	N° total de horas de atención al mes	CSA %[f]	Nº de usuarios que desistieron de la atención al mes	Nº total de usuarios atendidos al mes	DAP %[g]	Nº de llamadas no finalizadas por el usuario	Nº total de llamadas atendidas	CAT %[h]	Reclamos	Bajas	Consultas	Altas
Count[a]	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44
Meanb	2.363636	6.5	13.14614	26560.28	0.000498	16570.84	593611.2	0.027897	165590.3	3668369	0.043076	0.768636	0.748864	0.813636	0.914545
Std[c]	1.080287	3.63126	46.11138	3062.655	0.001774	6985.731	84467.2	0.011106	101745.4	579079.9	0.024298	0.066178	0.112853	0.051539	0.029131
min	1	1	0	8529	0	4370	99252	0.01	8856	1677328	0.0028	0.56	0.47	0.71	0.85
25%	1	3	0	26431.88	0	10507.75	586170	0.02	85795	3330132	0.02625	0.72	0.685	0.78	0.9
50%	2	6.5	0	27155	0	15717.5	602275	0.025911	174128	3757574	0.049295	0.78	0.765	0.82	0.92
75%	3	10	0	27895.5	0	19607	620695	0.031402	222426	4059456	0.05666	0.81	0.8225	0.85	0.94
max	4	12	244.16	29276.5	0.0097	32598	690478	0.053691	347867	5340026	0.096	0.88	0.94	0.92	0.96

Fuente: Elaboración propia.

Nota:

[a] Count: Cantidad de datos.

[b] Mean: Valor medio.

[c] Std: Desviación estándar.

[d] **Año:** Correspondiente a un periodo de 12 meses siendo, año 1 desde setiembre del 2014 a agosto del 2015, año 2 desde setiembre del 2015 a agosto del 2016, año 3 desde setiembre del 2016 a agosto del 2017 y año 4 desde setiembre del 2017 a abril del 2018.

[e] **Mes:** Correspondiente a un periodo mensual.

[f] **CSA%**: Tasa de caída del sistema de atención.

[g] **DAP (%):** Deserción en atención presencial.

[h] **CAT (%):** "Corte de la atención telefónica.

4.3.3.2 Resultados con Machine Learning

Tabla 9 Resultados Entel – XGBoost.

			Precisión	Exactitud	Sensibilidad	Puntaje F1	Curva ROC	M. Confusión
Sin Imputacion (mean - media)	Sin RFE	Sin Shrinkage	0.87500000	0.88888889	0.91666667	0.88311688	0.91666667	[[3 0] [1 5]]
		Shrinkage L1	0.87500000	0.88888889	0.91666667	0.88311688	0.91666667	[[3 0] [1 5]]
		Shrinkage L2	0.87500000	0.88888889	0.91666667	0.88311688	0.91666667	[[3 0] [1 5]]
		Shrinkage L1+L2	0.87500000	0.88888889	0.91666667	0.88311688	0.91666667	[[3 0] [1 5]]
	Con RFE*	Sin Shrinkage	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[3 0] [0 6]]
		Shrinkage L1	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[3 0] [0 6]]
		Shrinkage L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[3 0] [0 6]]
		Shrinkage L1+L2	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	[[3 0] [0 6]]

Con Imputacion (mean - media)	Sin RFE	Sin Shrinkage	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L1	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L2	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L1+L2	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
	Con RFE**	Sin Shrinkage	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L1	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L2	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]
		Shrinkage L1+L2	0.85714286	0.77777778	0.75000000	0.75000000	0.75000000[[5 0] [2 2]]

Fuente: Elaboración propia

* Variables consideras 2: CSA%, Consultas.

** Variables consideras 13: AÑO, Mes, Número de horas sin sistema de atención al mes, Número total de horas de atención al mes, CSA%, Número de usuarios que desistieron de la atención al mes, Número total de usuarios atendidos al mes, DAP%, Número de llamadas no finalizadas por el usuario, CAT%, Bajas, Consultas, Altas.

4.3.2.3 Diagrama de calor de indicadores de atención ENTEL

El algoritmo RFE; es una técnica de preprocesado cuyo objetivo es la selección de variables. Este algoritmo nos permite reducir variables eliminando las que tienen escasa y nula relación con el objetivo del estudio y cuando una o más variables contribuyen con la misma información respecto al objetivo del estudio, en la siguiente figura podemos observar la relación entre las variables.

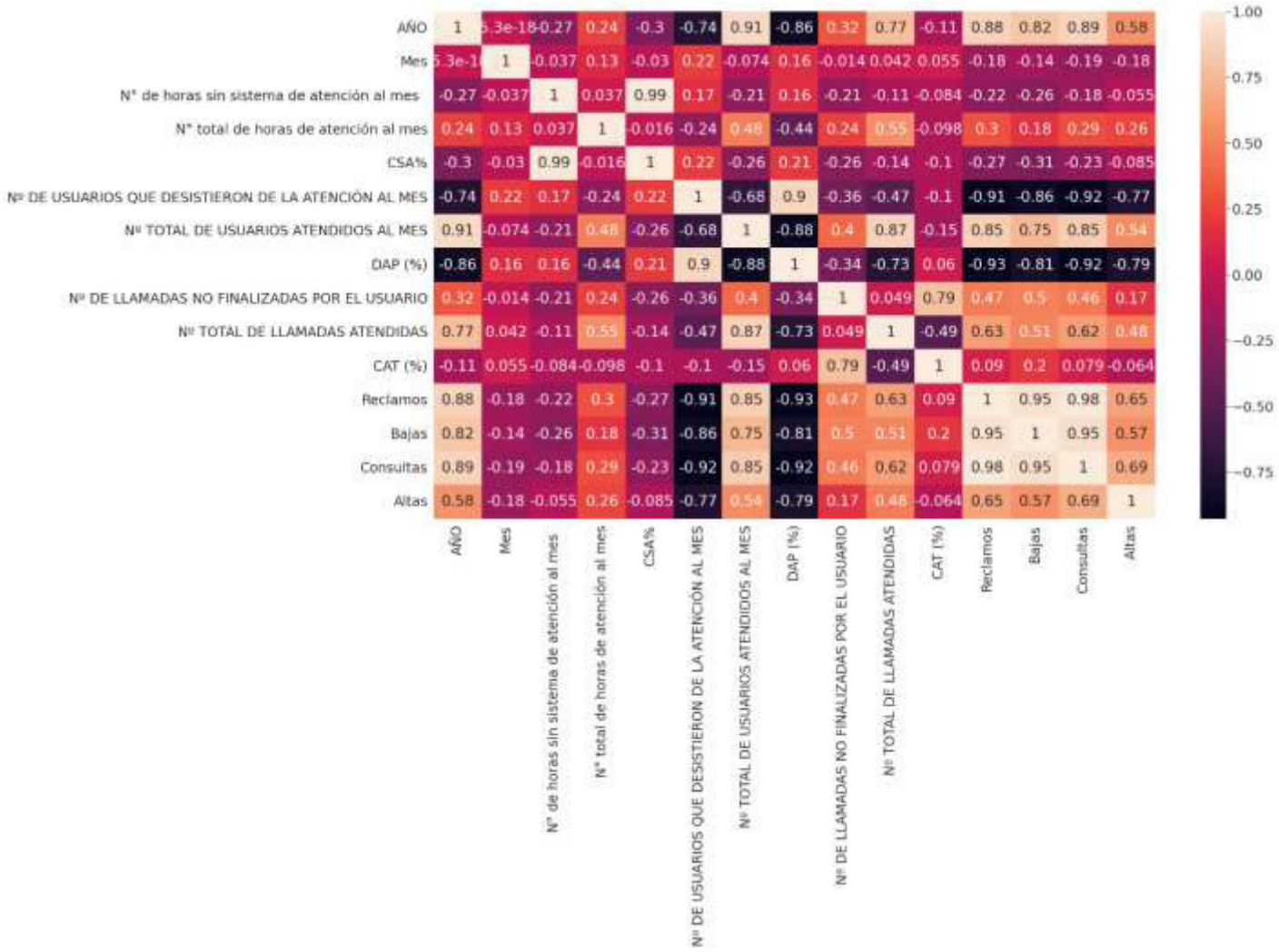

Figura 12 Diagrama de calor de indicadores de atención ENTEL

Fuente*:* Elaboración Propia.

4.3.2.4 Resultados comparativos ENTEL

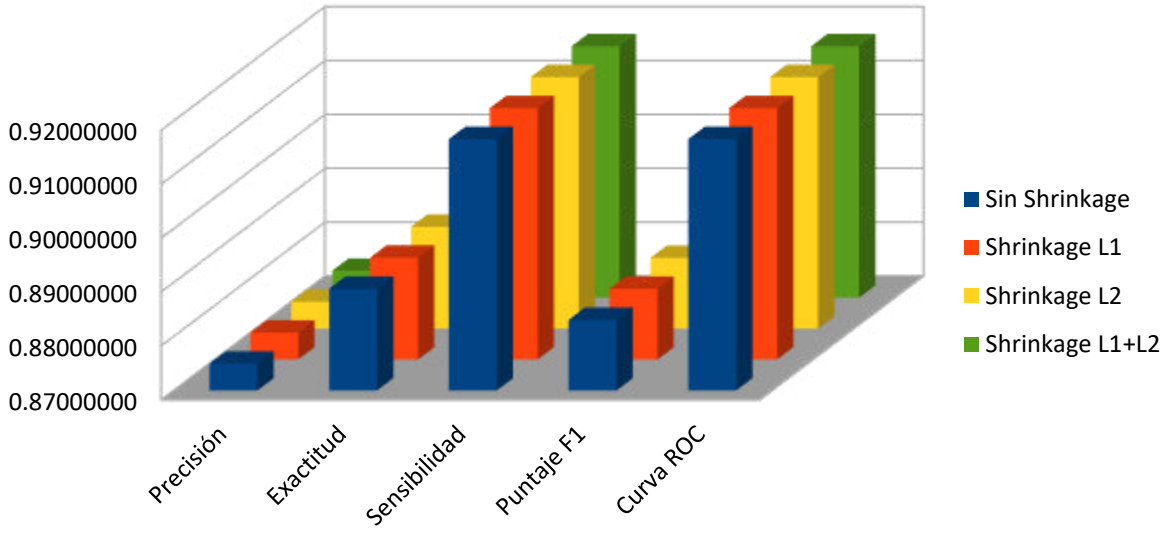

Figura 13 ENTEL Sin Imputación - Sin RFE.
Fuente*:* Elaboración Propia.

En la figura 13 podemos observar los resultados para el modelo sin imputación y sin RFE los cuatro algoritmos: Sin Shrinkage, Shrinkage L1, Shrinkage L2, Shrinkage L1+L2 obtuvieron una precisión de 87.5%, una exactitud de 88.89%, una sensibilidad de 91.67 y un puntaje F1 de 88.31%

Este modelo trabajó con las 15 variables debido a que no se utilizó ninguna técnica de RFE.

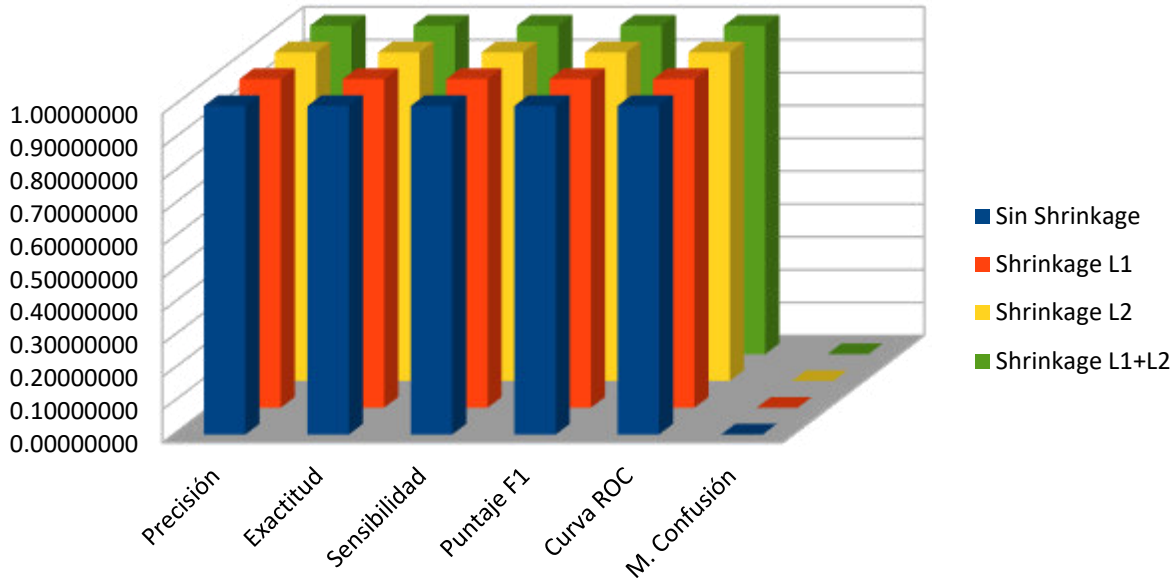

Figura 14 ENTEL Sin Imputación - Con RFE.

Fuente*:* Elaboración Propia.

En la figura 14 podemos observar los resultados para el modelo sin imputación y con RFE, en esta notamos que todos los modelos tienen Overfitting que es un concepto en Machine Learning que se da cuando el modelo propuesto coincide con los resultados de su entrenamiento. Cuando esto sucede, el modelo no se puede utilizar ya que habría memorizado.

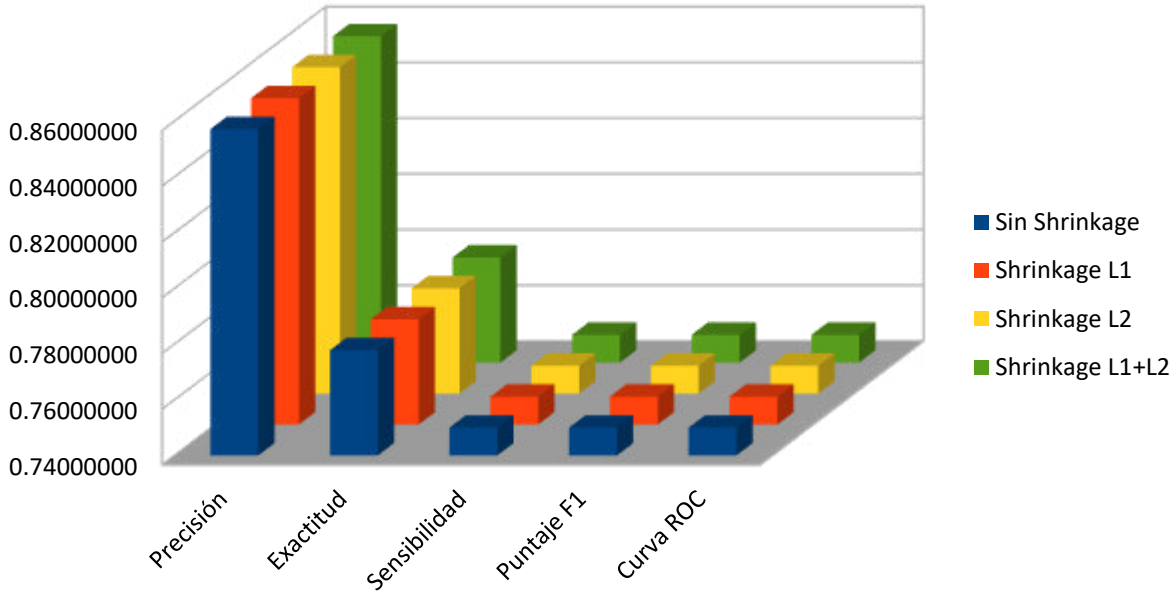

Figura 15 ENTEL Con Imputación - Sin RFE

Fuente*:* Elaboración Propia.

En la figura 15, se realizó imputación pues los datos no estaban completos sin el método RFE y los resultados obtenidos muestran que los cuatro algoritmos Sin Shrinkage, Shrinkage L1, Shrinkage L2, Shrinkage L1+L2 obtuvieron una precisión de 85.71%, una exactitud de 77.78%, una sensibilidad de 75.00% y un puntaje F1 de 75.00%.

Este modelo trabajó con las 15 variables debido a que no se utilizó ninguna técnica de RFE.

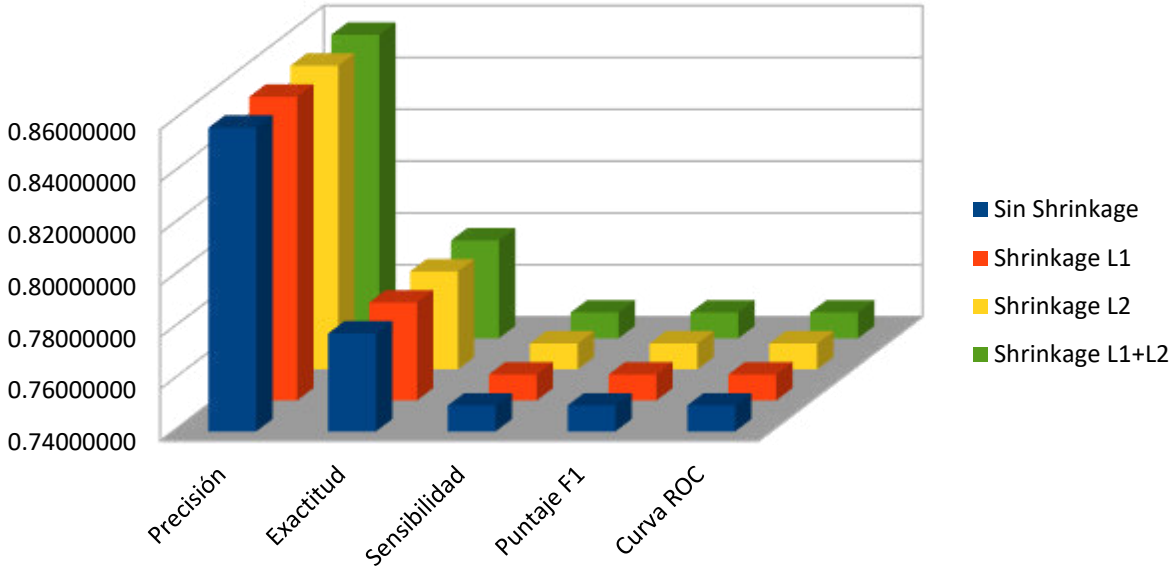

Figura 16 ENTEL Con Imputación - Con RFE

Fuente*:* Elaboración Propia.

En la figura 16, se realizó imputación pues los datos no estaban completos además se usó el método RFE y los resultados obtenidos muestran que los cuatro algoritmos Sin Shrinkage, Shrinkage L1, Shrinkage L2, Shrinkage L1+L2 obtuvieron una precisión de 85.71%, una exactitud de 77.78%, una sensibilidad de 75.00% y un puntaje F1 de 75.00%.

. También debo mencionar que después de aplicar el método hibrido con RFE tomando en cuenta solo valores con significancia real el modelo mantiene las siguientes variables independientes: AÑO, Mes, Número de horas sin sistema de atención al mes, Número total de horas de atención al mes, CSA%, Número de usuarios que desistieron de la atención al mes, Número total de usuarios atendidos al mes, DAP %, Nº de llamadas no finalizadas por el usuario, CAT %, Bajas, Consultas, Altas.

CONCLUSIONES

Según los objetivos propuestos se diferenció clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin por operador móvil arribando a las siguientes conclusiones.

- Movistar:

Se desarrolló modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación, estos fueron L1, L2, L1+L2 y Sin shrinkage; sin ninguna técnica de imputación de datos pues los valores proporcionados por OSIPTEL estaban completos, el modelo que mejor precisión obtuvo fue el Sin Shrinkage con 83.33% seguido de L1, L2 y L1+L2 todos con 67.85% de igual manera Sin Shrinkage fue el que mayor sensibilidad obtuvo con 92.85% seguido de L1, L2 y L1+L2 todos con 67.85%; concluyéndose que el modelo Sin Shrinkage fue el de mayor éxito.

También se desarrolló modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables RFE y extreme gradient boostin. El RFE redujo las variables de 15 a 7 considerando las siguientes ‘AÑO', 'CSA%', 'DAP%', 'CAT%', 'Bajas', 'Consultas', 'Altas': obteniéndose que RFE+L1 logro el mejor score con 93.75%, seguido de RFE+L2 y RFE+ Sin Shrinkage con 83.33% por último RFE+L1+L2 con 100% calificando como overfitting; concluyéndose que el modelo RFE+L1 fue el de mayor éxito.

- Claro:

Se desarrollo modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación, estos fueron L1, L2, L1+L2 y Sin shrinkage; con técnicas de imputación de datos pues los valores proporcionados por OSIPTEL estaban incompletos y los resultados presentaban overfitting, aplicando técnicas de imputación de datos los resultados fueron los siguientes, una precisión de 92.85% para L1, L2, L1+L2 y Sin shrinkage y una sensibilidad de 83.33% para L1, L2, L1+L2 y Sin shrinkage, concluyéndose que los 4 modelos obtuvieron los mismos puntajes.

También se desarrolló modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables RFE y extreme gradient boostin con técnicas de imputación de datos, El RFE redujo las variables de 15 a 8 considerando las siguientes 'Mes, N° de horas sin sistema de atención al mes, CSA %, DAP %, CAT %, Bajas, Consultas, Altas; se obtuvo los siguientes resultados L1, L2, L1+L2 y Sin shrinkage obtuvieron un score con 92.85%, de la misma manera todos los modelos obtuvieron una sensibilidad de 83.33% concluyéndose que los 4 modelos obtuvieron los mismo puntajes.

- Entel:

Se desarrollo modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación, estos fueron L1, L2, L1+L2 y Sin shrinkage; con técnicas de imputación de datos pues los valores proporcionados por OSIPTEL estaban incompletos y aplicándoles RFE presentaban overfitting, por lo cual se aplicó técnicas de imputación de datos los resultados fueron los siguientes, una precisión de 85.71% para L1, L2, L1+L2 y Sin shrinkage y una sensibilidad de 75.00% para L1, L2, L1+L2 y Sin shrinkage, concluyéndose que los 4 modelos obtuvieron los mismos puntajes.

También se desarrolló modelos de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables RFE y extreme gradient boostin con técnicas de imputación de datos, El RFE redujo las variables de 15 a 13 considerando las siguientes: Año, Mes, N° de horas sin sistema de atención al mes, Número total de horas de atención al mes, CSA%, Número de usuarios que desistieron de la atención al mes, Número total de usuarios atendidos al mes, DAP%, Número de llamadas no finalizadas por el usuario, CAT%, Bajas, Consultas, Altas; se llegó a los siguientes resultados L1, L2, L1+L2 y Sin shrinkage obtuvieron un score con 85.71%, de la misma manera todos los modelos obtuvieron una sensibilidad de 75.00% concluyéndose que los 4 modelos obtuvieron los mismo puntajes.

RECOMENDACIONES

- Sugerir que Osiptel publique mayor cantidad de datos (números y variables) para que los modelos propuestos puedan mejorar su entrenamiento y así poder proporcionar resultados con mayor precisión y exactitud.

- A los lectores e interesados en data science se recomienda obtener mayor cantidad de datos para aplicar otras técnicas de imputación como por ejemplo KNN.

REFERENCIAS BIBLIOGRÁFICAS

Albahli, Saleh, Muhammad Shiraz, and Nasir Ayub. 2020. “Electricity Price Forecasting for Cloud Computing Using an Enhanced Machine Learning Model.” *IEEE Access* 8:200971–81. doi: 10.1109/ACCESS.2020.3035328.

Andreas, Müller, and Guido Sarah. 2017. *Introduction to Machine Learning with Python: A Guide for Beginners in Data Science*.

Barupal, Dinesh Kumar, and Oliver Fiehn. 2019. “Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach.” *Environmental Health Perspectives* 127(9):2825–30. doi: 10.1289/EHP4713.

Baştanlar, Yalin, and Mustafa Özuysal. 2014. *Introduction to Machine Learning*. Vol. 1107. Cambridge University press.

Bergstra, James, Dan Yamins, and David Cox. 2013. “Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms.” *Proceedings of the 12th Python in Science Conference* (Scipy):13–19. doi: 10.25080/majora-8b375195-003.

Cajachahua Espinoza, Luis Angel. 2015. “Predicción de Fuga de Clientes : Una Aplicación de Técnicas de Data Mining En Telefonía Móvil.” Universidad Complutense de Madrid.

Chang, Yung Chia, Kuei Hu Chang, and Guan Jhih Wu. 2018. “Application of EXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions.” *Applied Soft Computing*

Journal 73:914–20. doi: 10.1016/j.asoc.2018.09.029.

Chen, Tianqi, Tong He, and Michael Benesty. 2018. “XGBoost : EXtreme Gradient Boosting.” *R Package Version 0.71-2* 1–4.

Deng, Yongshi, and Thomas Lumley. 2021. “Multiple Imputation Through XGBoost.” (2012).

Dua, Dheeru and Graff, Casey. 2017. “UCI Machine Learning Repository.” Retrieved September 6, 2021 (http://archive.ics.uci.edu/ml).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. “Additive Logistic Regression.” *The Annals of Statistics* 28(2):337–74.

Gonzales, Ligdi. 2013. “Machine Learning Con PYTHON Aprendizaje Supervisado.” *Journal of Chemical Information and Modeling* 53(9):1689–99.

Google. 2021. “Colaboratory.” Retrieved October 11, 2021 (https://colab.research.google.com/?utm_source=scs-index#scrollTo=5fCEDCU_qrC0).

Hartley, James. 2014. “Some Thoughts on Likert-Type Scales.” *International Journal of Clinical and Health Psychology* 14(1):83–86. doi: 10.1016/S1697-2600(14)70040-7.

Hou, Jiahui, Jianwei Qian, Yu Wang, Xiang Yang Li, Haohua Du, and Linlin Chen. 2019. “ML Defense: Against Prediction API Threats in Cloud-Based Machine Learning Service.” *Proceedings of the International Symposium on Quality of Service, IWQoS 2019*. doi: 10.1145/3326285.3329042.

Huaquipaco, Saul, Jose Cruz, Norman Jesus Beltran Castañon, Ferdinand Pineda, Christian Romero, Julio Fredy Chura Acero, and Wilson Mamani Machaca. 2021. “Modeling And Prediction Of A Multivariate Photovoltaic System, Using The Multiparametric Regression Model With Shrinkage Regularization And Extreme Gradient Boosting.” doi: 10.18687/laccei2021.1.1.557.

Khan, Prince Waqas, and Yung Cheol Byun. 2020. “Genetic Algorithm Based Optimized Feature Engineering and Hybrid Machine Learning for Effective Energy Consumption Prediction.” *IEEE Access* 8:196274–86. doi: 10.1109/ACCESS.2020.3034101.

Lo, Shuchuan. 2008. “Web Service Quality Control Based on Text Mining Using Support Vector Machine.” *Expert Systems with Applications* 34(1):603–10. doi: 10.1016/j.eswa.2006.09.026.

Martínez-Camblor, Pablo. 2007. “Comparación de Pruebas Diagnósticas Desde La Curva ROC.” *Revista Colombiana de Estadistica* 30(2):163–76.

Masso, Mauro Di, and Pablo M. Granitto. 2014. “Selección Estable de Variables Independientes Con RFE.” 26–34.

Medina, Fernando, and Marco Galván. 2007. *Imputación de Datos: Teoría y Práctica*. Vol. 4.

Mellado Ochoa, Abel Luis. 2013. “Análisis Sobre La Necesidad de Regular La Calidad Del Servicio de Telefonía Móvil En El Perú.” Pontificia Universidad Católica del Perú.

Mohana, R. S., and P. Thangaraj. 2013. “Machine Learning Approaches in Improving Service Level Agreement-Based Admission Control for a Software-as-a-Service Provider in Cloud.” *Journal of Computer Science* 9(10):1283–94. doi: 10.3844/jcssp.2013.1283.1294.

Muñoz Rosas, Francisco Juan, and Encarnación Álvarez Verdejo. 2009. “Métodos de Imputación Para El Tratamiento de Datos Faltantes: Aplicación Mediante R/Splus.” *Revista de Metodos Cuantitativos Para La Economia y La Empresa* 7(7):3–30.

Nakajima, Shin. 2018. “Quality Assurance of Machine Learning Software.” *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)* 601–4.

OSIPTEL. 2020. *Memoria Institucional OSIPTEL 2020*.

OSIPTEL. 2021. *Atención de Calidad*. Lima.

Rodrigo, Joaquín Amat. 2016. “Selección de Predictores y Mejor Modelo Lineal Múltiple : Subset Selection , Ridge Regression , Lasso Regression y Dimension Reduction.”

Valenzuela Najar, Jean Deynis. 2018. “Big Data Análisis de Comportamiento de Consumos de Clientes.”

de Vito, Laurent. 2017. “LinXGBoost: Extension of XGBoost to Generalized Local Linear Models.”

Wade, Corey. 2020. *Hands-On Gradient Boosting with XGBoost and Scikit-Learn*.

Wu, Aimin;, Lyn; March, Xuanqi; Zheng, Jinfeng; Huang, Xiangyang; Wang, Jie; Zhao, Fiona; M.Blyth, Emma; Smith, Rachelle; Buchbinder, and Damian; Hoy. 2020. “A Hybrid Method with TOPSIS and Machine Learning Techniques for Sustainable Development of Green Hotels Considering Online Reviews.” *Nature* 388:1–14.

Zolotareva, Ekaterina. 2021. “Aiding Long-Term Investment Decisions with XGBoost Machine Learning Model.” 1–29.

Albahli, Saleh, Muhammad Shiraz, and Nasir Ayub. 2020. “Electricity Price Forecasting for Cloud Computing Using an Enhanced Machine Learning Model.” *IEEE Access* 8:200971–81. doi: 10.1109/ACCESS.2020.3035328.

Andreas, Müller, and Guido Sarah. 2017. *Introduction to Machine Learning with Python: A Guide for Beginners in Data Science*.

Barupal, Dinesh Kumar, and Oliver Fiehn. 2019. “Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach.” *Environmental Health Perspectives* 127(9):2825–30. doi: 10.1289/EHP4713.

Başţanlar, Yalin, and Mustafa Özuysal. 2014. *Introduction to Machine Learning*. Vol. 1107. Cambridge University press.

Bergstra, James, Dan Yamins, and David Cox. 2013. “Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms.” *Proceedings of the 12th Python in Science Conference* (Scipy):13–19. doi: 10.25080/majora-8b375195-003.

Cajachahua Espinoza, Luis Angel. 2015. “Predicción de Fuga de Clientes : Una Aplicación de Técnicas de Data Mining En Telefonía Móvil.” Universidad Complutense de Madrid.

Chang, Yung Chia, Kuei Hu Chang, and Guan Jhih Wu. 2018. “Application of EXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions.” *Applied Soft Computing Journal* 73:914–20. doi: 10.1016/j.asoc.2018.09.029.

Chen, Tianqi, Tong He, and Michael Benesty. 2018. “XGBoost : EXtreme Gradient Boosting.” *R Package Version 0.71-2* 1–4.

Deng, Yongshi, and Thomas Lumley. 2021. “Multiple Imputation Through XGBoost.” (2012).

Dua, Dheeru and Graff, Casey. 2017. “UCI Machine Learning Repository.” Retrieved September 6, 2021 (http://archive.ics.uci.edu/ml).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. “Additive Logistic Regression.” *The Annals of Statistics* 28(2):337–74.

Gonzales, Ligdi. 2013. “Machine Learning Con PYTHON Aprendizaje Supervisado.” *Journal of Chemical Information and Modeling* 53(9):1689–99.

Google. 2021. “Colaboratory.” Retrieved October 11, 2021 (https://colab.research.google.com/?utm_source=scs-index#scrollTo=5fCEDCU_qrC0).

Hartley, James. 2014. “Some Thoughts on Likert-Type Scales.” *International*

Journal of Clinical and Health Psychology 14(1):83–86. doi: 10.1016/S1697-2600(14)70040-7.

Hou, Jiahui, Jianwei Qian, Yu Wang, Xiang Yang Li, Haohua Du, and Linlin Chen. 2019. "ML Defense: Against Prediction API Threats in Cloud-Based Machine Learning Service." *Proceedings of the International Symposium on Quality of Service, IWQoS 2019*. doi: 10.1145/3326285.3329042.

Huaquipaco, Saul, Jose Cruz, Norman Jesus Beltran Castañon, Ferdinand Pineda, Christian Romero, Julio Fredy Chura Acero, and Wilson Mamani Machaca. 2021. "Modeling And Prediction Of A Multivariate Photovoltaic System, Using The Multiparametric Regression Model With Shrinkage Regularization And Extreme Gradient Boosting." doi: 10.18687/laccei2021.1.1.557.

Khan, Prince Waqas, and Yung Cheol Byun. 2020. "Genetic Algorithm Based Optimized Feature Engineering and Hybrid Machine Learning for Effective Energy Consumption Prediction." *IEEE Access* 8:196274–86. doi: 10.1109/ACCESS.2020.3034101.

Lo, Shuchuan. 2008. "Web Service Quality Control Based on Text Mining Using Support Vector Machine." *Expert Systems with Applications* 34(1):603–10. doi: 10.1016/j.eswa.2006.09.026.

Martínez-Camblor, Pablo. 2007. "Comparación de Pruebas Diagnósticas Desde La Curva ROC." *Revista Colombiana de Estadistica* 30(2):163–76.

Masso, Mauro Di, and Pablo M. Granitto. 2014. "Selección Estable de Variables Independientes Con RFE." 26–34.

Medina, Fernando, and Marco Galván. 2007. *Imputación de Datos: Teoría y Práctica*. Vol. 4.

Mellado Ochoa, Abel Luis. 2013. “Análisis Sobre La Necesidad de Regular La Calidad Del Servicio de Telefonía Móvil En El Perú.” Pontificia Universidad Católica del Perú.

Mohana, R. S., and P. Thangaraj. 2013. “Machine Learning Approaches in Improving Service Level Agreement-Based Admission Control for a Software-as-a-Service Provider in Cloud.” *Journal of Computer Science* 9(10):1283–94. doi: 10.3844/jcssp.2013.1283.1294.

Muñoz Rosas, Francisco Juan, and Encarnación Álvarez Verdejo. 2009. “Métodos de Imputación Para El Tratamiento de Datos Faltantes: Aplicación Mediante R/Splus.” *Revista de Metodos Cuantitativos Para La Economia y La Empresa* 7(7):3–30.

Nakajima, Shin. 2018. “Quality Assurance of Machine Learning Software.” *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)* 601–4.

OSIPTEL. 2020. *Memoria Institucional OSIPTEL 2020*.

OSIPTEL. 2021. *Atención de Calidad*. Lima.

Rodrigo, Joaquín Amat. 2016. “Selección de Predictores y Mejor Modelo Lineal Múltiple : Subset Selection , Ridge Regression , Lasso Regression y Dimension Reduction.”

Valenzuela Najar, Jean Deynis. 2018. “Big Data Análisis de Comportamiento de Consumos de Clientes.”

de Vito, Laurent. 2017. “LinXGBoost: Extension of XGBoost to Generalized Local Linear Models.”

Wade, Corey. 2020. *Hands-On Gradient Boosting with XGBoost and Scikit-Learn*.

Wu, Aimin;, Lyn; March, Xuanqi; Zheng, Jinfeng; Huang, Xiangyang; Wang, Jie; Zhao, Fiona; M.Blyth, Emma; Smith, Rachelle; Buchbinder, and Damian; Hoy. 2020. “A Hybrid Method with TOPSIS and Machine Learning Techniques for Sustainable Development of Green Hotels Considering Online Reviews.” *Nature* 388:1–14.

Zolotareva, Ekaterina. 2021. “Aiding Long-Term Investment Decisions with XGBoost Machine Learning Model.” 1–29.

ANEXOS

Características Computacionales

Hardware:

Architecture: x86_64

CPU op-mode(s): 32-bit, 64-bit

Byte Order: Little Endian

CPU(s): 2

On-line CPU(s) list: 0,1

Thread(s) per core: 2

Core(s) per socket: 1

Socket(s): 1

NUMA node(s): 1

Vendor ID: GenuineIntel

CPU family: 6

Model: 79

Model name: Intel(R) Xeon(R) CPU @ 2.20GHz

Ram: 12.69 GB

Disco: 107.72 GB

CPU MHz: 2199.998

BogoMIPS: 4399.99

Hypervisor vendor: KVM

Virtualization type: full

L1d cache: 32K

L1i cache: 32K

L2 cache: 256K

L3 cache: 56320K

NUMA node0 CPU(s): 0,1

Software

DISTRIB_ID=Ubuntu

DISTRIB_RELEASE=18.04

DISTRIB_CODENAME=bionic

DISTRIB_DESCRIPTION="Ubuntu 18.04.5 LTS"

NAME="Ubuntu"

VERSION="18.04.5 LTS (Bionic Beaver)"

ID=ubuntu

ID_LIKE=debian

PRETTY_NAME="Ubuntu 18.04.5 LTS"

Versión de Python: 3.7.12

Matriz de consistencia

Modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin.

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES
Problema General	**Objetivo General**	**Hipótesis General**	**Dependiente**
¿Como influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin?	Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin	Es posible desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin	Calidad de servicio **Independientes** • Año • Mes • N° de horas sin sistema de atención al mes. • N° total de horas de atención al mes • Tasa de caída del sistema de atenciónCSA% • Nº de usuarios que desistieron de la atención al mes • Nº Total de usuarios atendidos al mes • Deserción en atención presencial DAP (%) • Nº de llamadas no finalizadas por el usuario • Nº total de llamadas atendidas • Corte de la atención telefónica CAT (%) • Reclamos • Bajas • Consultas. • Altas.
Problema Especifico	**Objetivos Específicos**	**Hipótesis Especificas**	
¿Cómo influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de clasificación?	Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de regression.	Es posible desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando técnicas de regression	
¿Cómo influye desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin?	Desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin	Se podrá desarrollar un modelo de clasificación y predicción multivariable de calidad de servicio mediante indicadores de atención utilizando métodos híbridos de selección de variables y extreme gradient boostin	