



**UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL**

**FACULTAD DE INGENIERÍA  
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**TEMA:**

**Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador**

**AUTOR:**

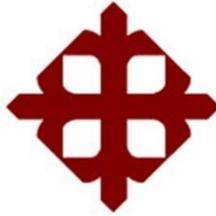
**Ronald David Solórzano Navarrete  
Trabajo de titulación previo a la obtención del grado de  
INGENIERO EN SISTEMAS COMPUTACIONALES**

**TUTOR:**

**Ing. Molina Flores, Gustavo, Mgs.**

**Guayaquil, Ecuador**

**22 de septiembre de 2020**



**UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL  
FACULTAD DE INGENIERÍA  
CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**CERTIFICACIÓN**

Certificamos que el presente trabajo de titulación **Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador**, fue realizado en su totalidad por **Ronald David Solórzano Navarrete** como requerimiento para la obtención del Título de **Ingeniero en Sistemas Computacionales**.

**TUTOR**

---

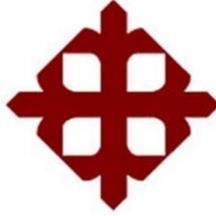
**Ing. Gustavo Molina Flores, Mgs.**

**DIRECTORA (e) DE LA CARRERA**

---

**Ing. Ana Isabel Camacho Coronel, Mgs.**

Guayaquil, a los 22 días del mes de septiembre del año 2020



**UNIVERSIDAD CATÓLICA DE  
SANTIAGO DE GUAYAQUIL**

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**DECLARACIÓN DE RESPONSABILIDAD**

Yo, Ronald David Solórzano Navarrete

**DECLARO QUE:**

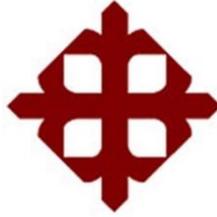
El Trabajo de Titulación **Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador**, previo a la obtención del Título de **Ingeniero en Sistemas Computacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 22 días del mes de septiembre del año 2020

**EL AUTOR**

---



**UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL  
FACULTAD DE INGENIERÍA  
CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**AUTORIZACIÓN**

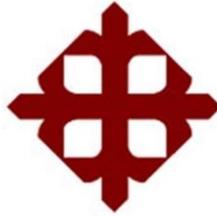
Yo, Ronald David Solórzano Navarrete

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación **Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 22 días del mes de septiembre del año 2020

**EL AUTOR**

---



UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES

REPORTE DE URKUND

URKUND

Document Information

Analyzed document	Tesis Ronald Solórzano [0009].docx (D78174304)
Submitted	8/26/2020 3:44:00 AM
Submitted by	
Submitter email	ronal.solorzano@cu.ucsg.edu.ec
Similarity	1%
Analysis address	gustavo.molina.ucsg@analysis.arkund.com

Sources included in the report

W	URL: <a href="https://link.springer.com/chapter/10.1007/978-3-030-30244-3_8#citeasGolava">https://link.springer.com/chapter/10.1007/978-3-030-30244-3_8#citeasGolava</a> Fetched: 8/26/2020 3:51:00 AM
W	URL: <a href="https://link.springer.com/content/pdf/10.1186/s12942-016-0056-6.pdf#m">https://link.springer.com/content/pdf/10.1186/s12942-016-0056-6.pdf#m</a> Fetched: 8/26/2020 3:51:00 AM

Fecha de elaboración: MARTES 25 DE Agosto, 2020

Firma:

Nombre del tutor: GUSTAVO MOLINA FLORES , MGE  
Tutor de Trabajo de Titulación  
Carrera de Sistemas Computacionales

TUTOR

f. \_\_\_\_\_

Ing. Gustavo Molina Flores, Mgs.

## **AGRADECIMIENTO**

Este trabajo se debe mucho a la colaboración de mis padres, a los cuáles le agradezco por su apoyo moral e interés, en especial por mi madre la Dr. Noemi Navarrete que siempre estuvo pendiente de mi crecimiento y bienestar a lo largo de la carrera ferviente para que pueda enfrentarme al mundo en las mejores condiciones y poder cumplir mis sueños, siempre será un pilar fundamental para mi crecimiento académico y mi vida.

## **DEDICATORIA**

Dedico este trabajo de titulación a mis padres, Rafael Solórzano y Noemi Navarrete, por su apoyo y amor inquebrantable a lo largo de este arduo y hermoso proceso, por siempre ver lo mejor de mí en todo momento, gracias por tener fe en mí y en mi capacidad para superar obstáculos y cumplir mis objetivos de vida.

# ÍNDICE GENERAL

<b>ÍNDICE GENERAL</b> .....	<b>VIII</b>
<b>ÍNDICE DE TABLAS</b> .....	<b>X</b>
<b>ÍNDICE DE FIGURAS</b> .....	<b>XI</b>
<b>RESUMEN</b> .....	<b>XII</b>
<b>ABSTRACT</b> .....	<b>XIII</b>
<b>INTRODUCCIÓN</b> .....	<b>2</b>
<b>CAPÍTULO I El Problema</b> .....	<b>4</b>
<b>Planteamiento del Problema</b> .....	<b>4</b>
Ubicación del problema en un contexto .....	4
Causas y consecuencias del problema .....	6
Delimitación del problema .....	7
Formulación del problema .....	7
Evaluación del problema .....	8
Objetivos .....	9
Objetivo General .....	9
Objetivos Específicos.....	9
Alcance del Problema .....	10
Justificación e Importancia .....	10
Pregunta de Investigación.....	11
Variables de la Investigación .....	11
<b>CAPÍTULO II Marco Teórico</b> .....	<b>12</b>
Inteligencia Artificial.....	12
Machine Learning .....	13
Interpretación en el ciclo de vida de la ciencia de datos.....	13
PDR características deseadas para interpretaciones .....	16
Interpretabilidad basada en modelos. ....	19
Interpretabilidad post hoc. ....	20
Sistemas de recomendación.....	21
La curva de Lorenz y el coeficiente de Gini .....	23
Métricas de similaridad para sistemas de recomendación .....	25
Distancia Euclidiana .....	26
Coeficiente de correlación de Pearson .....	26

Coseno de similaridad .....	27
<b>CAPÍTULO III Metodología de la investigación .....</b>	<b>29</b>
Tipo de investigación .....	29
POBLACIÓN Y MUESTRA .....	31
Población .....	31
Muestra .....	31
Instrumentos de recolección de datos .....	32
Instrumentos de la investigación.....	32
Análisis de los resultados de la entrevista.....	33
Metodología del sistema de recomendación.....	36
<b>CAPÍTULO IV .....</b>	<b>37</b>
<b>Propuesta .....</b>	<b>37</b>
Herramientas de desarrollo que se aplican para el diseño del sistema de recomendación .....	37
<b>Desarrollo del sistema de recomendación.....</b>	<b>39</b>
Problema, datos y audiencia. ....	39
Modelo (Sistema de recomendación).....	41
<b>Conclusión .....</b>	<b>51</b>
<b>Recomendaciones .....</b>	<b>52</b>
<b>ANEXOS.....</b>	<b>53</b>
Anexo 1 .....	53
Anexo 2 .....	53
Anexo 3 .....	54
Anexo 4 .....	54
Anexo 5 .....	55
<b>REFERENCIAS .....</b>	<b>56</b>

## ÍNDICE DE TABLAS

Tabla 1 Colaboradores entrevistados .....	34
Tabla 2 Guía de entrevista para cada colaborador .....	34
Tabla 3 Razones para usar Python.....	38
Tabla 4 Diccionario de datos. Elaboración propia .....	39

## ÍNDICE DE FIGURAS

<b>Figura 1</b> .....	14
<b>Figura 2</b> .....	19
<b>Figura 3</b> .....	24
<b>Figura 4</b> .....	26
<b>Figura 5</b> .....	27
<b>Figura 6</b> .....	27
<b>Figura 7</b> .....	28
<b>Figura 8</b> .....	29
<b>Figura 9</b> .....	30
<b>Figura 10</b> .....	40
<b>Figura 11</b> .....	41
<b>Figura 12</b> .....	42
<b>Figura 13</b> .....	43
<b>Figura 14</b> .....	44
<b>Figura 15</b> .....	45
<b>Figura 16</b> .....	46
<b>Figura 17</b> .....	46
<b>Figura 18</b> .....	47
<b>Figura 19</b> .....	48
<b>Figura 20</b> .....	48
<b>Figura 21</b> .....	49
<b>Figura 22</b> .....	50

## RESUMEN

La evolución tecnológica a lo largo de la historia de las empresas de telecomunicaciones ha generado nuevas formas de comunicar lo que ha llevado a generar servicios disruptivos en la industria. Hay que considerar la realidad actual donde hay un incremento abrumador del uso de datos por los usuarios y la aportación a la economía con el 5% en el PIB (en Latinoamérica), además del impacto positivo en empleos y servicios tecnológicos. Este trabajo “Diseñar un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador” tiene como objetivo recomendar paquetes prepagos a los usuarios de las diversas operadoras telefónicas. La metodología que se aplica en este trabajo es el ciclo de vida de la ciencia de datos para desarrollar el problema, modelo a utilizar y la métrica para obtener la precisión de este; sistemas de recomendación con enfoque de collaborative filtering; la curva de Lorenz y el coeficiente de Gini para determinar la distribución de los ingresos de los clientes. La herramienta de Jupyter Notebook y el lenguaje de programación Python se utiliza para aplicar la metodología del sistema propuesto el cual está compuesto por el análisis estadístico de los datos obtenidos, las variables primordiales generadas de las entrevistas con colaboradores de las empresas de telecomunicaciones, el algoritmo de recomendación de paquetes para cada usuario y la evaluación de la precisión de los resultados. Los resultados obtenidos por el sistema permiten concluir que el enfoque de collaborative filtering, las variables utilizadas en el sistema y el uso de la métrica de similitud de coseno permite realizar la recomendación de los paquetes a los usuarios con una precisión mayor al 60%.

**Palabras Clave:** *Machine Learning, Sistema de Recomendación, métodos estadísticos, coseno de similitud, Python, Collaborative Filtering.*

## ABSTRACT

Technological evolution throughout the history of telecommunications companies has generated new ways of communicating, which has led to the generation of disruptive services in the industry. We must consider the current reality where there is an overwhelming increase in the use of data by users and the contribution to the economy with 5% of GDP (in Latin America), in addition to the positive impact on jobs and technological services. This work "Designing a system of recommendations for prepaid packages using artificial intelligence techniques in telephone networks in Ecuador" aims to recommend prepaid packages to users of the various telephone operators. The methodology applied in this work is the life cycle of data science to develop the problem, model to be used and the metric evaluated to obtain its precision rating; recommendation systems with a collaborative filtering approach; the Lorenz curve and the Gini coefficient to determine the distribution of customer income. The Jupyter Notebook tool and the Python programming language are used to apply the methodology of the proposed system which is composed of the statistical analysis of the data obtained, the primary variables generated from the interviews with collaborators of the telecommunications companies, the algorithm recommendation of packages for each user and the evaluation of the accuracy of the results. The results obtained by the system led to the conclusion that the collaborative filtering approach, the variables used in the system, and the use of the cosine similarity metric allow the recommendation of the packages to the users with a precision greater than 60 %.

**Keywords:** *Machine Learning, Sistema de Recomendación, statistical methods, cosine similarity, Python, Collaborative Filtering.*

# INTRODUCCIÓN

Las empresas de telecomunicaciones a lo largo de la historia han evolucionado tecnológicamente lo que ha permitido mejorar la calidad y oferta de los servicios ya existentes, generar nuevas formas de comunicación y nuevos servicios disruptivos, tiene como consecuencia la expansión de esta industria hacia nuevos mercados como es la monetización de los datos.

Este mercado nace en base a la necesidad de explotar el incremento sustancial de datos que utilizan los usuarios por año, 5.8 Gb (Gigabytes) al año (2018) en Latinoamérica, con incremento proyectado del 392% para el año 2024 (O'Dea, 2020), lo que genera una gran cantidad de datos sobre el usuario en el tiempo, lo que permite demostrar un comportamiento. Esto se vuelve de gran valor para las empresas de telecomunicaciones, por esta razón es imprescindible el uso de técnicas de inteligencia artificial como machine learning, deep learning, redes neuronales artificiales, entre otras, para el tratamiento y explotación de estas grandes cantidades de datos.

Estas técnicas de inteligencia artificial permiten el descubrimiento de nuevos patrones claves en los datos como: las predicciones de compras de los usuarios a través de algoritmos de regresión, bayesianos, de árbol de decisión entre otros; mejorar la calidad de la red mediante la predicción de futuros problemas de conexión en la red de las empresas de telecomunicaciones además de casos de usos adicionales que permitirán reducir costos o incrementar los ingresos de las compañías.

A pesar de todo, el conjunto de datos que se encuentra en las empresas de telecomunicaciones no es apta para el consumo de las técnicas de inteligencia artificial ya que esta puede presentar muchos escenarios, como incongruencia en los tipos de datos observados, alta proporción de cantidad de nulos en los registros, duplicidad de registros, entre otros. Afectando de manera negativa a cualquier técnica de inteligencia artificial que se desee aplicar.

Debido a lo mencionado anteriormente, el presente trabajo de titulación presentará un sistema de recomendación de paquetes mediante el uso de técnicas de inteligencia artificial, el cual será sometido a un proceso de tratamiento de los datos presentados para incrementar su precisión. El trabajo presentado se desarrolla de la siguiente manera: Capítulo I refleja el problema, objetivos, alcance, justificación e importancia; Capítulo II presenta el marco teórico y toda la literatura que se utiliza en este trabajo de titulación; Capítulo III se describe la metodología a utilizar para este documento; Capítulo IV se describe el sistema de recomendación diseñado, con un detalle de los procesos realizados y el flujograma a seguir del algoritmo. Al final se presentarán las conclusiones y recomendaciones de la presente línea de investigación.

# **CAPÍTULO I**

## **El Problema**

Este capítulo hace referencia al problema de la investigación en las empresas de telecomunicaciones, como estas son afectadas, los objetivos que se plantean para resolverlo, el alcance de este trabajo y el sistema que se propondrá para solucionar el problema observado.

### **Planteamiento del Problema**

#### **Ubicación del problema en un contexto**

De acuerdo con Satari (2020), el uso de los datos móviles en Latinoamérica se ha incrementado en un 100% (2019 vs 2018), llegando a un promedio mensual de 4.7Gb por suscriptor en el 2020. Se espera que este crecimiento llegue hasta cinco veces más para el 2025, es importante considerar, Latinoamérica tiene la participación regular en aplicaciones digitales más alta de la región y es clave para el crecimiento del tráfico de datos. Al final del 2020 se observa una tasa de adopción de smartphones en la región de alrededor del 70%, lo que se traduce en una tendencia potenciada para el uso de datos móviles de los usuarios. Debido a la pandemia que se vive actualmente, una mayor cantidad de personas se ha visto obligada a quedarse en casa, lo que ha generado un incremento en el uso de tráfico de datos en un 30%, si se compara con los datos previos a la pandemia.

Las tecnologías móviles y servicios generaron un PIB del 5% en Latinoamérica en el 2018, esto representa una cantidad aproximada de \$260 billones de dólares de valor económico agregado. Además, la generación de estas tecnologías y servicios soportó 1,7 millones de trabajos (directa e indirectamente) y permitió recolectar más de \$38 billones de dólares mediante impuestos para el sector público (GSMA, 2019).

A medida que los países de la región se benefician incrementalmente por la mayor captación de servicios móviles y las mejoras asociadas en productividad y eficiencia se estima que para el año 2023, la contribución móvil

a la economía de Latinoamérica alcance los \$300 billones de dólares (GSMA, 2019).

La tendencia principal que ha surgido en las empresas de telecomunicaciones es la inteligencia artificial, por lo que éstas deben prepararse ante la emergente tecnología disruptiva, la cual se proyecta contribuir con quince trillones de dólares en la economía global para el 2030. Actualmente varios países en Latinoamérica se encuentran formulando estrategias y framework para soportar el desarrollo de inteligencia artificial y la implementación en la región (GSMA, 2019).

La creciente convergencia de mercados y productos, el constante cambio de modelos de negocio y el aumento de expectativas del cliente han creado un nivel de complejidad invisible para los proveedores de servicios de telecomunicaciones. En las industrias de telecomunicaciones la complejidad se ha segmentado en tres importantes grupos estos son el producto, procesos y tecnologías de la información (TI), basados en el hecho de que los catálogos de los productos se han vuelto mucho más diversos y, por lo tanto, deben ser respaldados por un proceso, históricamente complejo, y un “system landscape” (por ejemplo una plataforma como SAP) (Golavatchev, Buddle, & Hong, 2010).

En Ecuador desde el 2019 se evidencia una guerra de ofertas de servicios móviles entre las empresas de telecomunicaciones. Se materializa con el lanzamiento de planes prepagos disruptivos por parte de la Corporación Nacional de Telecomunicaciones (CNT) ya que ofrece planes con gran cantidad de datos móviles a un precio menor o similar que la competencia (dependiendo de la tarifa). Esto genera que las demás empresas de telecomunicaciones empiecen una fuerte campaña de planes prepago lo que incrementa la oferta y diversidad del catálogo de servicios disponibles para el cliente, como resultado este se vuelve complejo y diverso, lo que evita que el usuario final tome una decisión acertada, siendo altamente influenciado por los asesores comerciales, que éstas sesgados por su ganancia económica personal. Además, el marketing publicitario que se considera para estas campañas tiene como objetivo resolver las necesidades de las diversas

compañías y no la realidad del cliente, lo que se traduce en una gestión ineficiente (Orozco, 2019).

Debido a lo mencionado anteriormente, la industria de telecomunicaciones tiene la necesidad de realizar acciones disruptivas, respecto a la oferta del catálogo de servicios al cliente final. Se ha vuelto crítico para los proveedores de servicios, generar un catálogo de ofertas con paquetes atractivos y personalizados para cada cliente que permita solventar la necesidad real de este, permitiendo generar un valor agregado al servicio ofrecido, por esto se requiere aplicar un sistema de recomendación de paquetes mediante técnicas de inteligencia artificial.

### **Causas y consecuencias del problema**

En base a las entrevistas que se realizaron con colaboradores de Conecel S.A., existe una gran variedad de ofertas comerciales respecto a paquetes prepago en las empresas de telecomunicaciones, ocasionando confusión al cliente en el momento de intentar comprar un paquete. Se generan campañas enfocadas en los ingresos de la compañía, sin considerar la situación real del cliente como la situación económica que vive o las necesidades que tiene, éstas pueden ser priorizar datos sobre minutos de llamada o tal vez priorizar mensajes de texto sobre minutos y datos, todo esto y más conlleva a clientes insatisfechos con el producto adquirido. La falta de un sistema de recomendación de paquetes en base al requerimiento del cliente y la compañía provoca situaciones dónde el cliente se cambia a paquetes de menor valor en dólares lo que impacta de manera negativa a los ingresos de las compañías.

## Delimitación del problema

El problema se delimita en los siguientes campos:

Campo:	Inteligencia artificial
Área:	Empresas de telecomunicaciones del Ecuador
Aspecto:	Sistema de recomendación de paquetes prepago con enfoque de filtro colaborativo.
Tema:	Diseño de un sistema de recomendación usando Python en Jupyter Notebook que permita recomendar a un cliente un paquete prepago que se adapte las necesidades de este.

## Formulación del problema

La competitividad constante en el ámbito de las empresas de telecomunicaciones, la evolución tecnológica presente en la historia, la gran variedad de ofertas comerciales de paquetes prepago y la necesidad de ofertas personalizadas por el cliente ha impulsado a las diversas compañías a innovar en la forma que hacen negocios.

Esto se puede observar en, según el Comercio (2019), en el lanzamiento de la oferta prepago Plan Todos Conectados, de la CORPORACION NACIONAL DE TELECOMUNICACIONES CNT SOCIEDAD ANONIMA, con el precio final de \$7,84 (incluido el IVA) e incluye 2.000 megas de navegación libres, 200 megas adicionales por navegación para Facebook y 50 minutos gratis para a teléfonos fijos y móviles de la compañía. En su momento generó una reacción por parte de OTECEL S.A. y CONSORCIO ECUATORIANO DE TELECOMUNICACIONES S.A. CONECEL ofreciendo mejores servicios o reduciendo precios en su oferta comercial. Esto tiene

como consecuencia la competencia por la captación de abonados y últimamente afectando la decisión del cliente por la diversa variedad de ofertas comerciales y la insistencia de los asesores comerciales de las compañías.

En base a la problemática mencionada se formula el siguiente problema:

¿Cuáles son las variables necesarias que se deben considerar al momento del diseño del sistema de recomendación de paquetes prepago para responder a la necesidad real del cliente?

### **Evaluación del problema**

La ausencia de ofertas comerciales personalizables para cada usuario y la gran diversidad de paquetes prepago en las empresas de telecomunicaciones del Ecuador es un problema latente. La inteligencia artificial en la analítica de negocios se ha vuelto una carrera, y todas las diversas compañías quieren ganar la carrera, ser primeros en llegar al mercado con sus servicios, lo que conlleva a una alta competitividad entre las mismas. Esta usa como recurso principal los datos que se generan y recolectan en las empresas, debido a la pandemia el uso de datos de los usuarios ha incrementado por lo tanto la cantidad almacenada de los mismos también aumenta, esto refuerza la necesidad de utilizar técnicas de inteligencia artificial para su tratamiento y explotación.

Es relevante porque se basa en una nueva tecnología disruptiva que está emergiendo en el mercado de las empresas de telecomunicaciones de Latinoamérica, se espera que la inteligencia artificial contribuya a la economía global con \$15 trillones de dólares para el 2030 (GSMA, 2019).

Es evidente porque la inteligencia artificial se ha vuelto una necesidad en el mercado de telecomunicaciones y debido a la alta competitividad entre compañías para mantenerse a la vanguardia, es necesario la implementación

de esta en las estrategias de marketing que realizan para incrementar la tasa de conversión de las ventas y responder a las necesidades reales del cliente.

Es original porque luego de las entrevistas con los colaboradores de las empresas de telecomunicaciones al momento no existe una solución, como un sistema de recomendación de paquetes prepago mediante el uso de inteligencia artificial para los usuarios.

Es concreto ya que a pesar de que es aplicable a cualquier empresa de telecomunicaciones del Ecuador, puede ser personalizado y configurado para responder ante la realidad única de las diversas compañías.

Es factible ya que el recurso principal que se necesita para poder analizar la problemática, el conjunto de datos, ya se obtiene a partir del repositorio académico de Harvard y al ser una tecnología disruptiva existe información de calidad, aplicable a este problema, en todo el mundo.

## **Objetivos**

Los objetivos planteados están orientados a dar solución a la problemática presentada.

### **Objetivo General**

Diseñar un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador que permita la recomendación de paquetes a sus usuarios finales con una precisión mayor al 60%.

### **Objetivos Específicos**

Identificar los procesos que se involucran en el sistema de recomendación, con el fin de establecer las variables esenciales para el sistema.

Interpretar la distribución óptima de los datos mediante un análisis estadístico descriptivo para la transformación y limpieza de datos.

Aplicar técnicas de inteligencia artificial para diseñar el sistema de recomendación de paquetes aplicando lenguaje de programación Python en un Jupyter Notebook.

### **Alcance del Problema**

El presente proyecto diseñará un sistema de recomendación de paquetes prepago mediante técnicas de machine learning en un Jupyter Notebook, se basará en el conjunto de datos, por sus siglas en inglés, XDRs (X Data Record) la X representa cualquier registro ya sea de llamadas, datos, transacciones o sesiones. Además de datos demográficos de los suscriptores como sería en una de las empresas de telecomunicaciones del Ecuador.

Las técnicas de inteligencia artificial que se van a utilizar dependerán del análisis estadístico de los datos, pertinentes al dataset que se utilizará para este proyecto, en base a los insights obtenidos de dicho análisis permitirá determinar de manera sistémica cuál es la metodología más adecuada para el sistema de recomendación a desarrollar.

Este proyecto está categorizado dentro de la línea de investigación y desarrollo de nuevos servicios o productos de la carrera de Ingeniería de Sistemas Computacionales de la Facultad de Ingeniería de la Universidad Católica de Santiago de Guayaquil.

### **Justificación e Importancia**

En la actualidad, la inteligencia artificial se encuentra en auge en las diversas compañías de telecomunicaciones, por ser una tecnología disruptiva con un alto potencial académico, económico y social.

En las empresas de telecomunicaciones la inteligencia artificial permitirá interpretar patrones de gustos o de compra de sus suscriptores lo que permitirá generar campañas efectivas y costo-eficientes, además de tener una mayor tasa de conversión en sus ventas por su personalización y segmentación de sus abonados. Del lado del cliente permitirá generar ofertas que respondan a sus necesidades reales, sea por su capacidad económica o su tráfico en la red móvil.

En este trabajo de titulación, el resultado permite observar un sistema de recomendación de paquetes que ofrece de manera efectiva los paquetes vigentes en las empresas de telecomunicaciones a los suscriptores, contribuyendo a la fácil toma de decisión del cliente al momento de realizar la compra de un paquete entre todo el catálogo de oferta.

### **Pregunta de Investigación**

En base al trabajo de titulación presentado se busca evaluar y confirmar la siguiente pregunta de investigación planteada:

¿Es posible que un sistema de recomendación de paquetes prepago diseñado mediante técnicas de inteligencia artificial recomiende paquetes prepagos que cumplan con las necesidades reales de consumo del cliente?

### **Variables de la Investigación**

- **Variable independiente:** sistema de recomendación de paquetes prepago.
- **Variable dependiente:** recomendación de paquetes prepago mediante técnicas de inteligencia artificial.

# **CAPÍTULO II**

## **Marco Teórico**

### **Inteligencia Artificial**

A lo largo de la historia de la inteligencia artificial ha sido difícil llegar a un acuerdo respecto a su definición, sin embargo, luego de una extensa investigación se presenta las definiciones con mayor aceptación en la comunidad científica.

Según N. Kok (2009), la Inteligencia artificial, un área de estudio en el campo de las ciencias de la computación, está comprendida en el desarrollo de las computadoras que son capaces de participar en procesos del pensamiento similares a los humanos como aprendizaje, adaptabilidad, autocorrección, entre otros.

1. El concepto de que las máquinas pueden mejorar para asumir ciertas habilidades normalmente pensadas que son de la inteligencia humana como aprendizaje, adaptabilidad, autocorrección, entre otros.
2. Extensión de la inteligencia humana a través del uso de las computadoras, como en tiempos pasados el poder físico se extendía a través de las herramientas mecánicas.
3. En un sentido restringido, es el estudio de técnicas que usan las computadoras de manera efectiva para mejorar las técnicas de programación.

La aplicación de las técnicas de inteligencia artificial (IA) empezó a utilizarse en 1988 en las empresas de telecomunicaciones, y el uso principal era el diagnóstico de equipos complejos en la modalidad off-line. Desde entonces las empresas de telecomunicaciones y la IA han mejorado significativamente (Jiayin Qi, 2007).

La aplicación de la IA se puede dividir en dos tipos, Expert Systems (ES) y Machine Learning (ML). ES logra desenvolverse en una base de

conocimientos que ciertas veces representan hechos y reglas si-entonces, pero carece la habilidad de aprender de manera autónoma de datos exógenos mientras que ML es capaz de tener autoaprendizaje a partir de nuevas observaciones y crear reglas propias para clasificaciones complejas, generando un mayor potencial de transformación que ES (Balmer, Stanford, & Schmindt, 2020).

## **Machine Learning**

En base a lo mencionado de forma introductoria en el texto anterior, machine learning ha recibido un alto nivel de atención por su capacidad de predecir complejos fenómenos en las diversas industrias, han destacado en medicina, creación de políticas, en el campo científico y el área de telecomunicaciones (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Estas interpretaciones pueden volverse confusas por lo que es necesario definir y categorizar dichas interpretaciones en, model-based (creación de modelos de machine learning customizados para cada caso de uso) y post hoc (aplicación de métodos para analizar el modelo luego del entrenamiento). Luego es importante presentar el framework Predictive, Descriptive, Relevant (PDR), este marco de trabajo consiste en la selección de tres dimensiones deseadas de datos para la evaluación y construcción de métodos interpretativos: exactitud predictiva, exactitud descriptiva y relevancia, evaluada por una audiencia humana (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Se define la interpretación de ML como el uso de los modelos de machine learning para la extracción de conocimiento relevante acerca de las relaciones de dominio contenidas dentro de la data (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

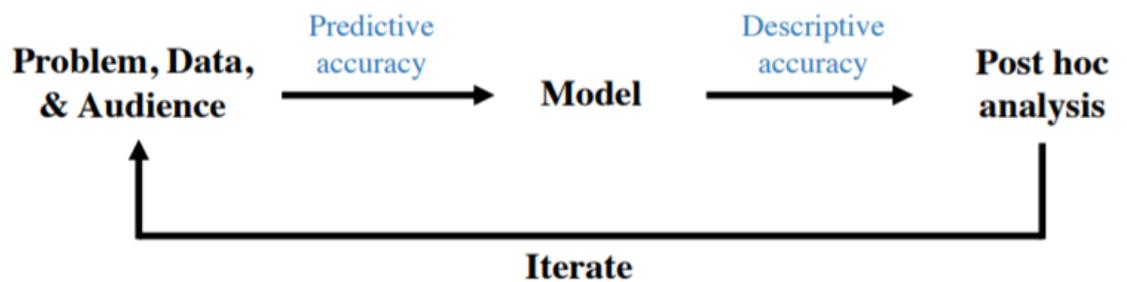
### **Interpretación en el ciclo de vida de la ciencia de datos**

Es importante explicar los procesos de la interpretación de ML dentro del ciclo de vida de la ciencia de datos. En la figura 1 se presenta de manera general una descripción de este proceso, con la intención de capturar la

mayoría de los problemas en la ciencia de datos. La interpretación en su gran parte ocurre en la fase de modelamiento y etapas de análisis post hoc, con el problema, la data y la audiencia proveen el contexto requerido para escoger el método apropiado (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

**Figura 1**

Resumen de las diferentes etapas (texto en negro) del ciclo de vida de la ciencia de datos.



Nota. Fuente tomada de (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)

Problema, Data, y Audiencia. Al inicio del ciclo, es necesario definir un problema de dominio que permita entender el uso de la data. Este problema puede tomar varias formas, desde un punto de vista científico, se puede enfocar en las relaciones contenidas en los datos, así como las células neuronales en un área en particular del sistema visual está relacionadas con el estímulo visual. En la perspectiva industrial, el problema concierne la capacidad de predicción u otras cualidades del modelo, como el asignar el score crediticio con una alta precisión, o hacerlo de manera distribuida y justa respecto al género y la raza. Nuevamente, se vuelve notable el rol importante de la interpretabilidad para la determinación del modelo adecuado a utilizar (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Luego de escoger el problema de dominio, es necesaria la recolección de data para el respectivo análisis. Aspectos de la recolección de los datos puede afectar el flujo de interpretación. Sesgo en los datos por problemas

comunes como desfase entre los datos recopilados y la población de interés se manifestarán en el modelo, restringiendo la habilidad de realizar interpretaciones respecto al problema de interés (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

El Modelo se basa en el problema escogido y la data recolectada, se construye un modelo predictivo. En este punto, se debe procesar, limpiar, y visualizar la data, extraer los features o variables, seleccionar el modelo (o varios modelos) y ajustarlos. La habilidad del modelo de ajustar la data es medida a través del acierto de predicción (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

El Análisis post hoc debe tener un modelo (o modelos) ajustados, permite realizar un análisis para obtener respuesta sobre la pregunta original. El proceso de analizar el modelo a veces involucra utilizar métodos de interpretabilidad para extraer varias formas de información (estables) del modelo. La información obtenida luego puede ser analizada y mostrada usando métodos estándares de analítica de datos, como diagramas de dispersión e histogramas (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Para la fase de Iterar, si las respuestas descubiertas luego del análisis post hoc son suficientes, se concluye el proceso. Caso contrario, se actualiza algún elemento del flujo (problema, data, y/o el modelo) y se itera el flujo mencionado, potencialmente varias veces (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Métodos de interpretación dentro del marco PDR. En el marco descrito anteriormente, la mayoría de los métodos de interpretación caen ya sea en las etapas de modelado o análisis post hoc. Se llama interpretabilidad en la etapa de modelado, esta parte de la capacidad de interpretación se centra en restringir la forma de los modelos ML para que puedan proporcionar fácilmente información sobre las relaciones descubiertas. Como resultado de estas limitaciones, el espacio de modelos potenciales es más pequeño, lo que puede resultar en una menor precisión predictiva. Por consiguiente, la

interpretabilidad basada en modelos se utiliza mejor cuando la relación subyacente es relativamente simple (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

### **PDR características deseadas para interpretaciones**

En general, no está claro cómo seleccionar y evaluar los métodos de interpretación para un problema y audiencia en particular. Para ayudar a guiar este proceso, se presenta el marco PDR, que consta de tres dimensiones que deben usarse para seleccionar métodos de interpretación para un problema particular: precisión predictiva, precisión descriptiva y relevancia (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Para la dimensión de precisión, La información que se genera por la interpretación de los métodos debe ser fiel al proceso subyacente que se desee tratar de entender. En el contexto de ML, hay dos áreas donde pueden surgir errores: al aproximar las relaciones de datos subyacentes con un modelo (precisión predictiva) y al aproximar lo que el modelo ha aprendido usando un método de precisión (precisión predictiva). Para que sea una interpretación confiable, se debe maximizar las precisiones en ambas áreas descritas (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Para la dimensión de precisión predictiva, la primera fuente de error ocurre durante la etapa de modelamiento, cuando un modelo de ML es construido el modelo aprende a partir de una pobre aproximación de una relación subyacente en los datos, cualquier información extraída del modelo es poco probable que sea acertada. Se debe evaluar si la calidad del modelo ajustado ha sido bien estudiada en frameworks estándares supervisados de ML, a través de medidas como la precisión del conjunto de prueba. En el contexto de interpretación, se describe este error como la precisión predictiva (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Cabe recalcar que este problema que involucra la interpretabilidad a veces requiere una noción de una precisión predictiva que vaya más allá de precisión promedio. La distribución de las predicciones es importante. Se puede dar el caso en que la predicción del error es muy alta para una clase

en particular lo que puede ser problemático a lo largo del proceso. Además, la precisión de la predicción debe ser estable respecto a datos razonables y perturbaciones del modelo. Por ejemplo, no se debe confirmar interpretaciones de un modelo que cambian dramáticamente cuando se entrena con un subconjunto de datos más pequeños de la data original (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Para la dimensión de precisión descriptiva. La segunda fuente de error ocurre durante la etapa de análisis post hoc, cuando el método de interpretación se usa para analizar el modelo ajustado. Ciertas veces, el método de interpretación provee una representación de las relaciones aprendidas por el modelo. Esto es especialmente desafiante para complejos modelos de caja negra como profundas redes neuronales, las cuales almacenan relaciones no lineales entre variables en formas no obvias (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Se define a la precisión descriptiva, en el contexto de interpretación, al grado en el cual el método de interpretación captura objetivamente las relaciones aprendidas por el modelo de machine learning (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Un conflicto común: precisión predictiva en contraste con la precisión descriptiva. Al momento de seleccionar qué modelo usar, los profesionales a menudo se enfrentan a una compensación entre la precisión predictiva y descriptiva. Por un lado, la simplicidad de los métodos de interpretación basada en modelos produce una precisión descriptiva consistentemente alta, pero a veces puede resultar en una menor precisión predictiva en conjuntos de datos complejos. Por otro lado, en entornos complejos como análisis de imagen, los modelos complicados generalmente proporcionan alta precisión predictiva, pero son más difíciles de analizar, lo que resulta en una menor precisión descriptiva (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

En la fase de relevancia, al seleccionar un método de interpretación, no solo este es suficiente para que el método tenga una alta precisión, la información extraída también debe ser relevante. Por ejemplo, en el contexto

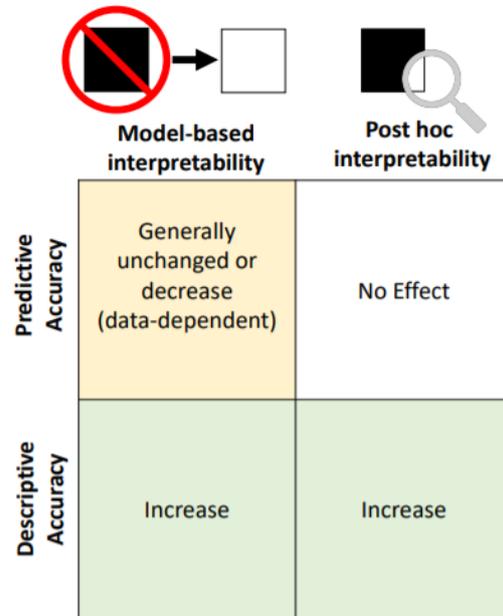
de la genética, un paciente, un médico, un biólogo y un estadístico pueden querer interpretaciones diferentes (pero consistentes) del mismo modelo. El contexto proporcionado por el problema, y las etapas de datos en la figura 1 guían qué tipos de relaciones se está interesado en aprender y, por extensión, los métodos que se deben utilizar.

Se define que una interpretación es relevante si esta provee insights para una particular audiencia dentro del problema de un dominio en específico. La relevancia a menudo juega un papel clave en la determinación del equilibrio entre la precisión predictiva y descriptiva. Dependiendo del contexto del problema, se puede elegir si se debe enfocar en uno sobre el otro (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Habiendo esbozado las principales características deseadas para la interpretación de métodos, se discute como se vincula con la interpretación en las etapas de modelado y análisis post hoc en el ciclo de vida de la ciencia de datos. La figura 2 dibuja paralelos entre las características deseadas para técnicas de interpretación y la categorización de métodos introducidos anteriormente. En particular, métodos post hoc y basados en modelos tienen como objetivo aumentar la descripción precisión, pero solo los métodos basados en modelos impactan la precisión predictiva. No se muestra la relevancia, que determina qué tipo de salida (output) es útil para un problema y audiencia en particular (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Figura 2

Impacto de los métodos de interpretabilidad en las precisiones descriptivas y predictivas.



Nota. Fuente tomada de (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)

### Interpretabilidad basada en modelos.

Se discute cómo participa las consideraciones de interpretabilidad en la etapa de modelado del ciclo de vida de la ciencia de datos (ver figura 1). En esta etapa, se construye un modelo de ML de los datos recopilados, se define la interpretabilidad basada en la construcción de esquemas que facilitan una idea de las relaciones que han aprendido. Diferentes métodos de interpretación basados en modelos proporcionan diferentes formas para aumentar la precisión descriptiva mediante la construcción de modelos que sean más fáciles de entender, a veces resulta en una menor precisión predictiva. El principal desafío de la interpretabilidad basada en modelos es generar modelos que sean lo suficientemente simples para ser fácilmente entendidos por el público, pero lo suficientemente sofisticado para ajustarse adecuadamente a los datos subyacentes (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Al seleccionar un modelo para resolver un problema de dominio, se debe considerar la totalidad del framework PDR. El primer deseo para considerar es la precisión predictiva. Si el modelo construido no representa con precisión el problema subyacente, cualquier análisis posterior será sospechoso. Segundo, el propósito principal de los métodos de interpretación basados en modelos es aumentar la precisión descriptiva (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Finalmente, la relevancia de la salida de un modelo debe considerarse y está determinada por el contexto del problema, los datos y la audiencia. Los métodos que son usados ampliamente son sparsity, simulatability, modularity, domain-based feature engineering y model-based feature engineering (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

### **Interpretabilidad post hoc.**

Se discute cómo entra en juego las consideraciones de interpretabilidad en la etapa de análisis post hoc del ciclo de vida de la ciencia de datos. En esta etapa, se analiza un modelo entrenado para proporcionar información sobre las relaciones aprendidas. Esto es particularmente desafiante cuando los parámetros del modelo no muestran claramente qué relaciones ha aprendido el modelo. Para ayudar en este proceso, una variedad de métodos de interpretación post hoc han sido desarrollados para proporcionar una idea de lo que un modelo entrenado ha aprendido, sin cambiar el modelo subyacente. Estos métodos son particularmente importantes para realizar configuraciones donde los datos recopilados son de alta dimensión y complejos, como es la data de imagen. Una vez que la información ha sido extraída del modelo ajustado, se puede analizar usando técnicas estándar exploratorias de análisis de datos, como diagramas de dispersión e histogramas (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

Al realizar un análisis post hoc, el modelo ya ha sido entrenado, por lo que su precisión predictiva es fija. Por lo tanto, bajo el framework PDR, solo se debe considerar la precisión descriptiva y la relevancia (en relación con un público en particular). Mejorar cada uno de estos criterios son áreas de una

investigación activa. Los métodos de interpretación post hoc más útiles caen en dos categorías principales: interpretaciones a nivel de predicción y a nivel de conjunto de datos, que a veces se denominan interpretaciones locales y globales, respectivamente. Los métodos de interpretación a nivel de predicción se centran en explicar las predicciones individuales realizadas por modelos, como qué características e interacciones llevaron a dicha predicción en particular. Los enfoques a nivel de conjunto de datos se centran en las relaciones globales que el modelo ha aprendido, como qué patrones visuales están asociados con una respuesta pronosticada. Estas dos categorías tienen mucho en común (de hecho, a nivel de conjunto de datos los enfoques a menudo producen información a nivel de predicción), pero se discuten por separado, ya que métodos en diferentes niveles son significativamente diferentes (W. Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019).

### **Sistemas de recomendación**

Los sistemas de recomendación son componentes de software que realizan personalizadas sugerencias a los usuarios, el cual se asume se ajusta a las preferencias del usuario. La idea general detrás de estos sistemas es de la explotación de información acerca del usuario, artículos y las relaciones entre ellos, así como ratings o acciones de compra de manera que permita identificar conjuntos fortuitos, por ejemplo, sugiriendo artículos a usuarios que de otra manera no los hubiera encontrado. Estos sistemas pueden ser aplicados a varios dominios, tales como sitios web de e-commerce, recomendación de películas o servicios turísticos, entre otros (Rossetti, Stella, & Zanker, 2013).

Los sistemas de recomendación pueden estar basados en dos tipos de estrategias. El content filtering (filtrado de contenido) se enfoca en la generación de perfilamientos para cada usuario o producto para caracterizar dicha naturaleza. Los perfiles de usuarios pueden incluir información demográfica o respuestas provistas en una encuesta o formulario. Los perfiles permiten a los programas asociar usuarios con productos que emparejen adecuadamente. Las estrategias de content filtering requieren la generación

de información externa que no puede estar disponible o sea fácil de recolectar (Koren, Bell, & Volinsky, 2009).

Una alternativa a content filtering, dónde solo se enfoca en el comportamiento pasado del usuario, por ejemplo, transacciones previas realizadas o rating (clasificación) de productos, sin la necesidad de la creación explícita de perfiles. Este enfoque se conoce como collaborative filtering (relleno colaborativo). Collaborative filtering analiza relaciones entre usuarios y sus interdependencias entre los productos para identificar nuevas asociaciones de usuario-artículo (Koren, Bell, & Volinsky, 2009).

El mayor atractivo de collaborative filtering es que de dominio libre, lo que permite tratar aspectos de los datos que a veces son elusivos y difíciles de perfilar usando content filtering. Aunque generalmente más preciso que técnicas basadas en content filtering, collaborative filtering sufre de un problema que se llama cold start (inicio frío) debido a su inhabilidad de tratar con los nuevos productos y usuarios del sistema. En este aspecto, content filtering es superior (Koren, Bell, & Volinsky, 2009).

Las dos principales áreas de collaborative filtering son neighborhood methods (métodos de vecindario) y latent factor models (modelos de factores latentes). Neighborhood methods está centrados en la relación computacional entre artículos o, alternativamente, entre usuarios. Este enfoque orientado a artículos evalúa las preferencias del usuario para un artículo basado en ratings de neighboring (vecino) artículos por el mismo usuario. Un producto vecino son el resto de los productos que tienden a obtener ratings similares cuándo son clasificados por el mismo usuario. Cabe recalcar que en caso de no existir ese rating de manera explícita puede ser generado de manera implícita (Koren, Bell, & Volinsky, 2009).

Latent factor models, también conocidos como Matrix Factorization (MF) models, son un enfoque alternativo que intenta explicar los ratings que caracterizan a ambos artículos y usuarios sobre el 20 a 100 factores que se pueden inferir de los patrones de ratings. Matrix factorization (matriz de factorización), en su forma básica, caracteriza ambos artículos y usuarios

mediante vectores de factores inferidos de patrones de clasificaciones de artículos. Una alta correlación entre el artículo y el factor del usuario conlleva a una recomendación. Estos modelos ofrecen una buena escalabilidad y precisión de predicción, además de la flexibilidad del modelo para adaptarse a los casos de usos reales que se presenten (Koren, Bell, & Volinsky, 2009).

Los sistemas de recomendación dependen de varios tipos de inputs de datos, el cuál es representado en una matriz de una sola dimensión mostrando los usuarios y otras dimensiones de los artículos de interés. Los datos que mejor ayudan como input a estos modelos son de tipo feedback (retroalimentación), explícito de alta calidad, lo que incluye un input explícito por los usuarios respecto a su interés en los productos. Por lo general, para el feedback explícito es necesario realizar una matriz dispersa, ya que es muy probable que cada usuario solo haya clasificado un pequeño porcentaje de los posibles productos (Koren, Bell, & Volinsky, 2009).

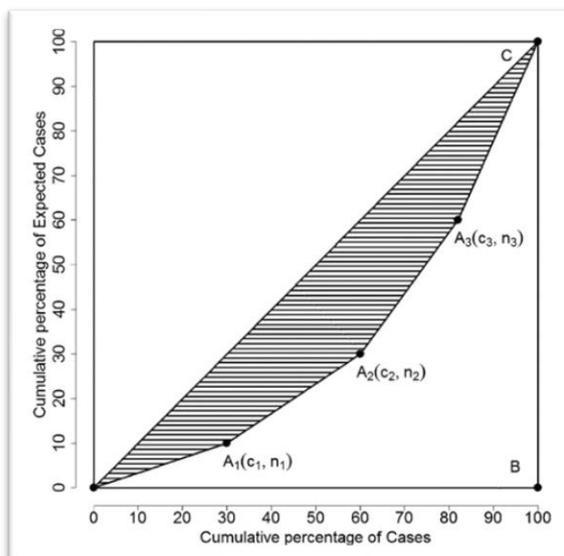
Una de las ventajas de la matriz de factorización es que permite la adición de información adicional. Cuando el feedback explícito no está disponible, los sistemas de recomendación pueden inferir las preferencias de los usuarios usando feedback implícito, lo cual indirectamente refleja la opinión mediante la observación del comportamiento del usuario, incluyendo historial de compra, historial de navegación, patrones de búsqueda, o incluso movimientos del mouse. Feedback implícito usualmente denota la presencia y ausencia de un evento, así que típicamente es representado por una matriz densamente llena (Koren, Bell, & Volinsky, 2009).

### **La curva de Lorenz y el coeficiente de Gini**

En economía, la curva de Lorenz usualmente es utilizada para explicar y medir la heterogeneidad de la distribución de la riqueza. Es una representación gráfica de la función acumulada de la distribución de la empírica probabilidad de la distribución. El formato básico es un rectángulo dividido entre dos triángulos rectángulos isósceles simétricos (como se denota en la figura 3).

Figura 3

Curva de Lorenz.



Nota. Tomado de (Han, y otros, 2016)

En esta figura, el punto  $A_i$  ( $c_i, n_i$ ) denota que el fondo  $n_i$  % de la población es dueña de  $c_i$  % de la riqueza total y el punto  $A_i$  está ordenado de manera que  $n_i$  se ordene de manera no decremental. La línea OC en el triángulo (línea  $y = x$ ) muestra una equidad de distribución de la riqueza y es usada como línea de referencia. Las dos piernas del triángulo, OB y OC, representan la perfecta inequidad de la distribución de la riqueza, por ejemplo, ninguna casa posee riquezas a excepción de la última que posee toda la riqueza de la sociedad. La curva OA1A2A3C dentro del triángulo se denomina curva de Lorenz y describe la observada, distribución de la riqueza. La curva de Lorenz es luego comparada contra la referencia de la línea OC (perfecta equidad de distribución de riquezas) (Han, y otros, 2016).

El coeficiente de Gini fue propuesto por un famoso economista italiano. Este coeficiente se basa en la curva de Lorenz, actualmente se utiliza internacionalmente como un indicador de análisis utilizado para realizar una investigación comprensiva en la interna diferencia de distribución de ingresos de los residentes (Wu, y otros, 2017).

La característica más predominante del coeficiente de Gini es la integridad, la cual permite mostrar la inequidad de ingresos de los residentes de una región como un todo. Además, por su integridad, cuándo una región tiene un número menor de residentes individuales, la efectividad del coeficiente de Gini es mayor. (Wu, y otros, 2017).

Estos conceptos mencionados son luego utilizados para la selección de características del dataset y la generación de clusters (grupos) para obtener una uniformidad de los datos.

### **Métricas de similitud para sistemas de recomendación**

La similitud entre dos usuarios es calculada con la ayuda de ratings hechos por otros usuarios. Encontrar la similitud entre estos usuarios es la tarea más crucial porque la precisión y la calidad de las recomendaciones depende en su gran mayoría de estas métricas. Existen varias métricas de similitud para encontrar la similitud entre los usuarios y productos lo que arroja un grado de cercanía y separación entre usuarios y productos. Escoger la métrica perfecta de similitud es esencial para los sistemas de recomendación con enfoque de collaborative filtering ya que diferentes métricas tendrán como resultado diferentes resultados en diferentes contextos de la información (Sondur, Chigadani, & Nayak, 2016).

Para poder implementar su función principal, se deben identificar los productos que sean útiles para el usuario, un sistema de recomendación debe predecir si un producto el cual vale la pena ser recomendado. Para realizar esto el sistema debe tener la habilidad predecir la utilidad de alguno de los productos o al menos comprar la utilidad de algunos de ellos y luego decidir qué producto recomendar en base a esta comparación. Es probable que el paso de predicción no sea explícito en el algoritmo de recomendación, sin embargo, este modelo unificado puede ser aplicado para describir el rol general de un sistema de recomendación (Sondur, Chigadani, & Nayak, 2016).

A continuación, se presentan las técnicas más populares que son usadas en sistemas de recomendación con enfoque collaborative filtering según (Sondur, Chigadani, & Nayak, 2016).

### **Distancia Euclidiana**

Las bases de varias métricas de similitud y disimilitud es la distancia euclidiana. Esta distancia entre vectores X y Y se define de la siguiente manera:

**Figura 4**

*Fórmula de la distancia euclidiana.*

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$

Nota. Tomada de (Sondur, Chigadani, & Nayak, 2016)

Se observa en la figura 4, dónde Xi y Yi son puntajes de clasificación de un elemento dada por dos usuarios diferentes para el mismo número de productos N, es el número de elementos comúnmente puntuados (Sondur, Chigadani, & Nayak, 2016).

Es decir, la distancia euclidiana es la raíz cuadrada de la suma de las diferencias al cuadrado entre los elementos correspondientes de los dos vectores. El coeficiente de correlación está (inversamente) relacionado con la distancia euclidiana entre las versiones estandarizadas de los datos (Sondur, Chigadani, & Nayak, 2016).

### **Coefficiente de correlación de Pearson**

A diferencia del puntaje de similitud de la distancia Euclidiana (escala entre 0 y 1), esta métrica evalúa que tan altamente correlacionadas se encuentran dos variables y es medido entre -1 y +1. Un coeficiente de

correlación de Pearson de 1 indica que los objetos de la data se encuentran perfectamente correlacionados, pero en este caso, un puntaje de -1 significa que los objetos de la data no se encuentran correlacionados. En otras palabras, el puntaje de la correlación de Pearson cuantifica como dos objetos de datos se ajustan en una línea (Sondur, Chigadani, & Nayak, 2016). En la figura 5 se observa la ecuación utilizada para obtener el coeficiente de correlación de Pearson.

**Figura 5**

*Fórmula de la correlación de Pearson.*

$$PC(u, v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)^2 \sum_{i \in I_{uv}} (r_{v,i} - \bar{r}_v)^2}}$$

*Nota. Tomada de (Sondur, Chigadani, & Nayak, 2016)*

### **Coseno de similitud**

Usualmente la métrica de coseno de similitud es usada para estimar la similitud entre dos instancias A y B en la recuperación de información que los objetos tienen forma de vectores Xa y Xb y calculando el coseno del vector (CV) (espacio del vector) similar entre los vectores, indica la distancia de ellos entre cada uno (Sondur, Chigadani, & Nayak, 2016).

**Figura 6**

*Fórmula del coseno de similitud.*

$$\frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

*Nota. Fuente tomada de (Sondur, Chigadani, & Nayak, 2016)*

Como se observa en la figura 6,  $A_i$  es el puntaje del usuario A y  $B_i$  es el puntaje del usuario B por el mismo elemento  $n$  es el número común de los elementos puntuados. En el contexto de los elementos recomendados, para el cómputo de usuarios similares, esta métrica puede ser empleada cuando un usuario  $u$  indica un vector  $x_u$ , dónde  $x_{ui} = r_{ui}$  si el usuario  $u$  ha puntuado el elemento  $i$  y para los que no hayan sido clasificados se los considera como 0 (Sondur, Chigadani, & Nayak, 2016). Esta similaridad entre dos usuarios  $u$  y  $v$  pueden ser calculados como en la figura 7.

**Figura 7**

*Fórmula del coseno de similaridad.*

$$CV(u, v) = \cos(X_a, X_b) = \frac{\sum_{i \in I_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2} \sqrt{\sum_{j \in I_v} r_{vj}^2}}$$

Nota. Fuente tomada de (Sondur, Chigadani, & Nayak, 2016)

Dónde  $I_{uv}$  una vez más indica los elementos puntuados por  $u$  y  $v$ . Una desventaja de esta métrica es que no examina las diferencias en la moda y la varianza de las puntuaciones hechas por los usuarios  $u$  y  $v$  (Sondur, Chigadani, & Nayak, 2016).

A partir de la observación realizada por Sondur, Chigadani y Nayak (2016) la métrica de similaridad de coseno es la que muestra mayor precisión de recomendación en los diversos escenarios planteados, por lo que esta es la métrica que se utiliza en el sistema de recomendación para este trabajo de titulación.



diferentes insights en base a las métricas que se requieran evaluar para cumplir el criterio de negocio objetivo. Además, es necesario el uso de algoritmos estadísticos, técnicas de machine learning (ML) para generar data con un comportamiento aleatorio, distribuido y normalizado, lo cual es esencial para la construcción de modelos, también mediante este enfoque permitirá encontrar patrones en los datos que agreguen valor en la interpretación de los resultados del sistema de recomendación de paquetes (Hernández, Fernández, & Del Pilar, 2014).

**Figura 9**

*Proceso Cualitativo.*



Nota. Fuente tomada de (Hernández, Fernández, & Del Pilar, 2014)

## **POBLACIÓN Y MUESTRA**

### **Población**

El presente trabajo de titulación está orientado para las empresas de telecomunicaciones del Ecuador, según lo revisado en la superintendencia de compañía, valores y seguros en el portal de información del sector societario existen cientos de compañías dentro de la categoría de telecomunicación, sin embargo, el trabajo de titulación se enfoca en la recomendación de paquetes prepago, segmento móvil de las telecomunicaciones, lo cual solo las siguientes compañías ofrecen este servicio: OTECEL S.A., CONSORCIO ECUATORIANO DE TELECOMUNICACIONES S.A. CONECEL y CORPORACION NACIONAL DE TELECOMUNICACIONES CNT SOCIEDAD ANONIMA. Por lo tanto, se puede descartar el resto de las compañías que proveen otro tipo de servicios.

### **Muestra**

La muestra que se obtuvo para este trabajo de titulación se basa en la entrevista de cuatro colaboradores, de las compañías mencionadas en el párrafo anterior, con roles importantes los cuales se encuentran intrínsecamente relacionados con el contexto de investigación además tienen el respaldo de los varios años de experiencia desempeñados en la empresa.

Cabe recalcar que el presente trabajo de titulación no utilizará a la entrevista como método de recolección de información, a razón de que la naturaleza de la investigación se basa en la revisión y análisis de los datos obtenidos en los registros de XDRs cifrados. Sin embargo, se utiliza la entrevista para validar los criterios de negocio que le sirven a las diversas compañías, determinar la necesidad del negocio y definir las variables que agreguen valor al sistema de recomendación para incrementar su precisión de recomendación a los usuarios finales.

## **Instrumentos de recolección de datos**

La técnica que se utilizó para la recolección de datos fueron entrevistas de tipo abiertas a los colaboradores de las empresas mencionadas ya que se requiere obtener información del perfilamiento del cliente, su patrón de compra en paquetes prepago y las necesidades de negocio de las compañías para este segmento.

Toda técnica o instrumento de recolección de datos debe contar con tres requisitos esenciales: confiabilidad, validez y objetividad. Determinar la confiabilidad en una técnica de recolección de datos se debe al grado que se obtiene por la aplicación repetida al mismo individuo u objeto que produce resultados idénticos (Hernández, Fernández, & Del Pilar, 2014). Para contar con la validez se debe obtener el grado cuando una técnica mide específicamente la variable que pretende medir (Hernández, Fernández, & Del Pilar, 2014).

Se complica obtener objetividad en las técnicas de recolección, sin embargo, se logra mediante el consenso o mediciones múltiples. Esta es capaz de ser reforzada mediante la evaluación de resultados y la generación de procesos estándares respecto a la práctica de la técnica utilizada (clausulas y pasos a seguir para todos los participantes) (Hernández, Fernández, & Del Pilar, 2014).

## **Instrumentos de la investigación**

La entrevista se enfoca en los colaboradores que tienen alguna relación con el tema de investigación del trabajo de titulación, sea el área de marketing dónde se generan los ofertas comerciales de paquetes prepago, la venta directa de estos paquetes o la atención al cliente para el cambio de paquetes prepago e incluso el área que se encarga de analizar grandes cantidades de datos para generar campañas de marketing que ayuden a mejorar la tasa de conversión de las ventas de paquetes prepago de las compañías.

Debido a las políticas internas de confidencialidad de cada empresa se ha decidido salvaguardar la información personal de los colaboradores y la

empresa dónde trabaja. Se debe recalcar que los diversos roles de los entrevistados aportan de manera positiva a que se realice una recolección de datos exitosa.

La entrevista tiene como objetivo la determinación de las siguientes características:

- Identificar si existen soluciones informáticas que usen técnicas de inteligencia artificial.
- ¿Son efectivas las campañas de marketing que priorizan los ingresos de la compañía sobre las necesidades del cliente?
- Determinar si los entrevistados tienen experiencia en la venta, realizan analítica de negocio o generan estrategias de marketing de paquetes de prepago en las empresas de telecomunicaciones.
- Descubrir variables o características que describan al cliente.
- Definir la importancia de la situación económica del cliente al momento de tomar una decisión para la compra de los paquetes prepago.

Es importante mencionar que las entrevistas, orientadas en un contexto nacional por los entrevistados, ayudaron a la generación de variables importantes que representen el comportamiento del cliente, así como, la filtración de variables que no agregan valor o incluso incrementan el porcentaje de error en el sistema de recomendación.

Debido a las políticas internas de confidencialidad de cada empresa se ha decidido salvaguardar la información personal de los colaboradores y la empresa dónde trabaja.

### **Análisis de los resultados de la entrevista**

Las entrevistas se realizaron a cuatro colaboradores, uno del área de marketing, otro del área de analítica de negocio y el dos del área del servicio de atención al cliente, estas se hicieron de manera digital y se entregó una guía de preguntas previamente a los entrevistados.

Tabla 1 Colaboradores entrevistados

ENTREVISTADO	TEMA TRATADO	CODIFICACION
Ingeniero comercial	Estrategia de MKT en clientes prepagos	C1
Ingeniero en administración de empresas con maestría en estadística	Técnicas estadísticas para la venta de paquetes prepago	C2
Asesor comercial de ventas	Venta de paquetes prepago	C3, C4

Nota. Se da un código a cada colaborador entrevistado (C). Elaboración propia

Debido a las diferentes funciones que desempeña cada uno de los colaboradores en las compañías, se realizaron preguntas diferentes en ciertos casos, dependiendo del conocimiento o especialidad de ellos.

Tabla 2 Guía de entrevista para cada colaborador

Pregunta	C1	C2	C3	C4
1	¿Existe una solución informática que aplique inteligencia artificial en la empresa?	¿Se utiliza técnicas de inteligencia artificial para la clusterización de clientes prepago en las campañas de marketing?	¿Existe una solución informática que aplique inteligencia artificial en la empresa?	¿Existe una solución informática que aplique inteligencia artificial en la empresa?
2	¿En base a que criterios actualmente se le recomienda un paquete prepago al cliente?	¿En base a que criterios actualmente se le recomienda un paquete prepago al cliente?	¿En base a que criterios actualmente se le recomienda un paquete prepago al cliente?	¿En base a que criterios actualmente se le recomienda un paquete prepago al cliente?
3	¿Cuál es el segmento de clientes que se encuentra más afectado a nivel de ingresos para la compañía al momento,	¿Cuál es el segmento de clientes que se encuentra más afectado a nivel de ingresos para la compañía al momento, prepago o pospago?	¿Cuál es el segmento de clientes que se encuentra más afectado a nivel de ingresos para la compañía al momento, prepago o pospago?	¿Cuál es el segmento de clientes que se encuentra más afectado a nivel de ingresos para la compañía al momento, prepago o pospago?

	prepago o pospago?			
4	¿Qué tipos de estrategia de MKT se utiliza para la venta de paquetes prepago?	¿Cuáles son las características más importantes para un cliente al momento de comprar un paquete prepago?	¿Cuáles son las características más importantes para un cliente al momento de comprar un paquete prepago?	¿Cuáles son las características más importantes para un cliente al momento de comprar un paquete prepago?

Nota. Elaboración propia

Las entrevistas realizadas a los colaboradores permiten llegar a las siguientes conclusiones:

- Las empresas donde trabajan no cuentan soluciones informáticas que utilicen técnicas de inteligencia artificial.
- Los criterios que se utilizan para recomendar un paquete varían, en el caso del área de marketing se basa en el comportamiento anterior de los clientes para definir estos criterios, en el caso del asesor comercial presenta recomendaciones en base al input que tenga del cliente u ofrece paquetes en promoción.
- El segmento de clientes que presenta mayores ingresos en el segmento de clientes de pospago debido al modelo de negocio que se maneja.
- Se pudo definir las características más importantes que se toman en consideración para los clientes al momento de comprar paquetes prepagos.
- Se identificó las estrategias de marketing clásicas que se utilizan para la oferta de paquetes prepago.

El resultado de estas entrevistas ayudó a la generación del sistema de recomendación, en el análisis de datos para determinar las variables más

importantes y eliminar aquellas que no agreguen valor al sistema. Además del enfoque que debe tener el sistema para poder responder las necesidades del cliente y la empresa.

### **Metodología del sistema de recomendación**

Para esta sección se da una pequeña descripción de la metodología propuesta, se explica los criterios sobre el cuál se generó el dataset (conjunto de datos) y se presentan los métodos a evaluar para el sistema.

Las empresas de telecomunicaciones se apoyan en campañas de marketing para incrementar sus ventas y fidelizar sus clientes, especialmente para incrementar el Average Revenue Per User (ingreso promedio por usuario, en adelante ARPU) y reducir su tasa de churn (tasa de cancelación de los clientes). Las medidas que puede tomar a través de su plataforma son él envió de notificaciones de las campañas a sus clientes. Estás notificaciones pueden ser enviadas mediante SMS, Interactive Voice Response (IVR) o correo electrónico (Alves, Valente, Filipe, Castro, & Macedo, 2019).

El tipo de campaña que los clientes reciben puede estar compuesta por incentivos de carácter informativo, recargas o consumos, las cuáles pueden estar segmentada por las características (sociodemográficas o financieras) de los clientes. Estas pueden tener diferentes objetivos (incrementar ventas, ofrecer paquetes de datos, paquetes de recargas, incrementar market share (cuota de mercado) en el producto prepago, fidelizar al cliente, entre otros). Las notificaciones deben ser enviada a los clientes en base a eventos que las habiliten, como cumplir ciertas características, por ejemplo, llegar a cierto valor de su paquete contratado o entrar a un área específica. Es necesario considerar lo mencionado como criterio de negocio para saber qué información es relevante al momento de construir el sistema de recomendación (Alves, Valente, Filipe, Castro, & Macedo, 2019).

## CAPÍTULO IV

### Propuesta

El capítulo IV presenta el desarrollo del sistema de recomendación de paquetes prepago mediante el uso de técnicas de inteligencia artificial. Se describe cada una de las etapas que se realizó dentro de la metodología utilizada para la elaboración del sistema, así como el resultado, es decir, el porcentaje de acierto y error de este.

#### **Herramientas de desarrollo que se aplican para el diseño del sistema de recomendación**

Para seleccionar el lenguaje de desarrollo y el ambiente de trabajo dónde iba a realizar el proyecto fue necesario revisar diferentes papers de investigación, foros de científico de datos y benchmark de su performance. En este caso se selecciona trabajar con Python como lenguaje de desarrollo y Jupyter Lab como entorno gráfico integrado (IDE).

De acuerdo con artículo publicado en la IEEE ha sido una tradición entre los científicos de datos y desarrolladores el uso de lenguajes como C++, C y LISP pero en años recientes la popularidad de Python ha incrementado y a lo que va de Julio 2020 según (Jansen, 2020) el índice de TIOBE (indicador de popularidad de lenguajes de programación) Python se encuentra como el 3er lenguaje de programación más usado en el mundo (Nagpal & Gabrani, 2019). Se enumeran las razones que son de gran importancia para este proyecto en la tabla 3.

Tabla 3 Razones para usar Python.

<b>Razones para usar Python</b>	<b>Descripción</b>
<b>Menos codificación y alta legibilidad</b>	Si se compara Python vs otros lenguajes, este requiere 1/5 de líneas de código para aplicar la misma lógica que otros lenguajes de programación orientado a objetos. Vea Anexo 1.
<b>Portabilidad y Flexibilidad</b>	Ofrece la opción de escoger un enfoque orientado a objetos o desarrollar mediante scripts. Además de permitir el enlace entre diferentes datos estructurados que pueden ser usado como backend language.
<b>Plataforma independiente</b>	Python ofrece su propio API (interfaz de aplicación de programas) junto a que es independiente de plataformas, es decir, desarrolladores pueden cambiar el código del proyecto con leves modificaciones para correr en varios sistemas operativos (SO).
<b>Balance de programación de bajo y alto nivel</b>	Python permite balancearse entre alto y bajo niveles de programación. Lo que permite obtener diferentes tiempos de ejecución, siendo los bajos niveles de programación los más eficientes. Vea Anexo 2.
<b>Estructuras de datos</b>	Es imprescindible para un proyecto que se use inteligencia de datos el uso correcto de datos estructurados para un algoritmo.
<b>Bibliotecas de código abierto disponibles</b>	Es un lenguaje open source y debido a esto tiene una gran cantidad de librerías que se utilizan en la mayoría de los proyectos de inteligencia artificial.

Nota. Elaboración propia. Información tomada de (Nagpal & Gabrani, 2019)

Se utiliza Jupyter Lab por su intuitiva e interactiva interfaz para ambientes de ciencia de datos. Además de servir como una herramienta de presentación del proyecto desarrollados. Según estadísticas realizadas a más de 1900 personas, (Piatetsky, 2018) determina que el IDE más utilizado es Jupyter.

# Desarrollo del sistema de recomendación

## Problema, datos y audiencia.

Es importante describir el conjunto de datos utilizados en el sistema de recomendación.

Los datos que se utilizan en este proyecto se obtuvieron del repositorio: “Dataverse” (repositorio de datos opensource) de la universidad de Harvard (2020), se obtuvo 1 dataset (conjunto de datos) con 1,23 millones de registros referentes a los datos del usuario y sobre la transaccionalidad de los paquetes. Sin embargo, es necesario el análisis y transformación de los para la generación de un producto final de datos que simulara la realidad de Ecuador por lo que se procedió a separar el mismo en dos visiones, uno que muestre solo los datos del usuario y otra de los paquetes comprados en el periodo observado. Se describe los campos utilizados en el siguiente diccionario de datos.

Tabla 4 Diccionario de datos. Elaboración propia

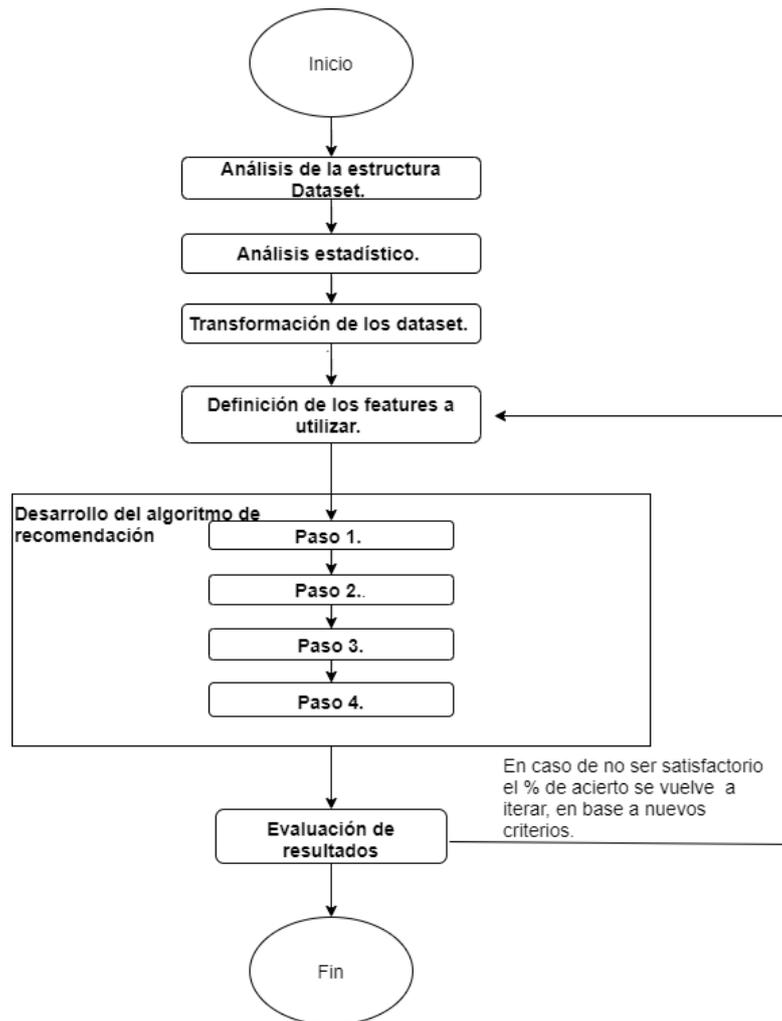
Variables	Tipo de dato	Descripción
MSISDN	int64	Número asociado a la línea del usuario.
EDAD	float64	Edad asociada al usuario
PROVINCIA	object	Estado asociado al usuario.
NSE	float64	Nivel socioeconómico al usuario
GENERO	object	Genero asociado al usuario.
GENERACION	object	Generación de edad asociada al usuario
CLT_INGRESOS	object	Clasificación según la cantidad de ingresos por paquetes para el periodo de 30 días asociada al usuario
FECHA_ACTIVACION	object	Fecha en la que se realizó la activación del paquete
MONTO	float64	Monto en dólares del paquete comprado.
PCK_TIPO	object	Tipo de paquete referente a la compra realizada.
GRUPO	object	Grupo de paquetes según su denominación en dólares.

La audiencia que se utiliza para resolver este problema es la muestra de los abonados con características representativas de las empresas de telecomunicaciones, no existe manipulación sobre el tipo de audiencia que se utiliza, es el reflejo directo de lo que se observa en los conjuntos de datos. Se realiza un muestreo aleatorio simple por conveniencia en función del acceso de los datos, se obtuvo una aproximación a los valores de la región.

Para poder describir el sistema de recomendación de paquetes de manera adecuada es importante mencionar el flujograma de los procesos que se involucran para el diseño de este, se irá detallando más adelante como funciona cada proceso de este.

**Figura 10**

*Flujograma general del sistema de recomendación de paquetes.*



Nota. Elaboración del autor.

## Modelo (Sistema de recomendación)

En base a la observación de los datos y su comportamiento, se obtuvo como insight que en los ingresos o consumos del cliente existen grupos que se alejan de los valores normales dentro del conjunto de datos, por ejemplo, se observa que la mayor cantidad de usuarios tiene consumos de paquetes de \$1 dólar, pero existen grupos muy pequeños, conocidos como valores atípicos, que consumen \$1.01, \$1.02, \$1.03, \$1.04, \$1.05 unidades de dólar lo que incrementa el porcentaje de error de la recomendación del sistema.

Para esto es necesario realizar un análisis estadístico que permita lograr una distribución normalizada de los datos, como se puede observar en la figura 11, se generan métricas estadísticas como: cantidad de registros, media, desviación estándar, mínimo, máximo y los percentiles lo que permite tener una idea de cómo se comportan los datos y su distribución.

Figura 11

*Análisis de las métricas estadísticas.*

	Unnamed: 0	EDAD	NSE	ING_CLTE	MSISDN
<b>count</b>	100000.000000	99253.000000	95483.000000	100000.000000	100000.000000
<b>mean</b>	49999.500000	40.510070	3.817947	2.418782	50000.470480
<b>std</b>	28867.657797	15.135775	0.596137	1.203544	28867.658325
<b>min</b>	0.000000	1.000000	1.000000	0.002004	1.000000
<b>25%</b>	24999.750000	29.000000	4.000000	1.000000	25000.750000
<b>50%</b>	49999.500000	38.000000	4.000000	2.000897	50000.500000
<b>75%</b>	74999.250000	49.000000	4.000000	3.078346	75000.250000
<b>max</b>	99999.000000	120.000000	5.000000	5.000000	100000.000000

Nota. Elaboración del autor.

Se investigó de varias fuentes para determinar la métrica estadística que permite resolver a la distribución inequívoca de estos respecto a los ingresos o consumos de un periodo del cliente.

Muchos autores para resolver este problema utilizan el índice de Gini, es una medida estadística que se utiliza para evaluar la inequidad económica, al ser la variable objetiva, los ingresos del cliente, este concepto puede ser aplicado en el sistema de recomendación para lograr el objetivo mencionado en el párrafo anterior.

**Figura 12**

*Código del análisis de estadístico.*

```
Analisis de los datos

[15]: ### Para evaluar el coeficiente Ginni
n_bins = 5

df_ginni = df.copy()

v = df_ginni['INGRESO_MES'].to_numpy(dtype=None, copy=False)
bins, result, gini_val = G(v, n_bins)
plt.figure()
plt.subplot(2, 1, 1)
plt.plot(bins, result, label="Observados")
plt.plot(bins, bins, '--', label="Perfecta ig.")
plt.xlabel("Fracción de población")
plt.ylabel("Fracción de riqueza")
plt.title("GINI: %.4f" %(gini_val))
plt.legend()
plt.subplot(2, 1, 2)
plt.hist(v, bins=n_bins)

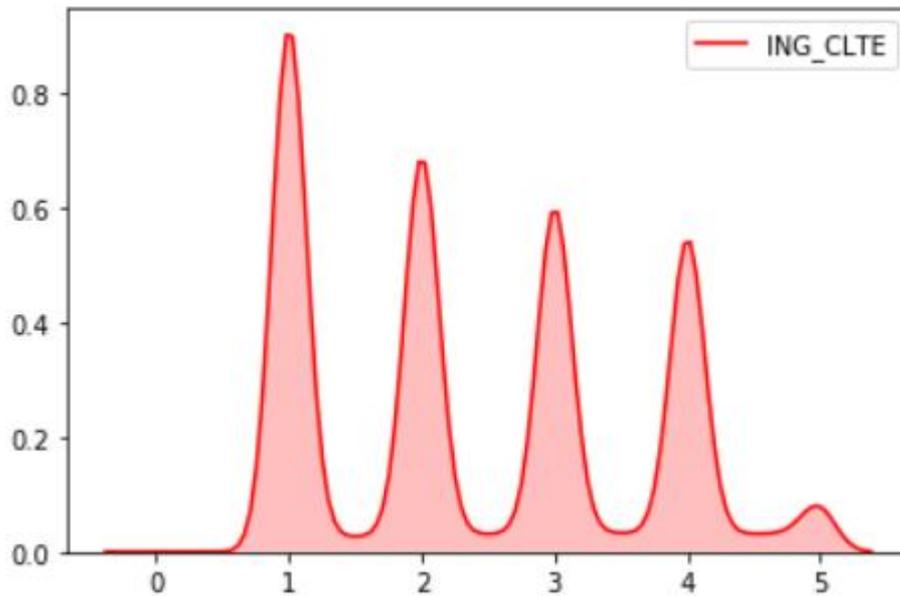
# Ver el % de frecuencia. value counts de todsos los cortes, sacarle el %.
```

Nota. Elaboración del autor.

Este es el código (figura 12) que se utiliza para realizar la evaluación Gini, donde se pasa como input en la función, la variable objetiva que se va a evaluar (ingresos del cliente) y la cantidad de bins (contenedores) en los cuáles se va a agregar los datos.

**Figura 13**

*Densidad de la variable de ingresos del cliente.*



Elaboración del autor.

Nota.

En base a los insights de la figura 11 y al observar la densidad de la variable de la figura 13, se puede determinar que el agrupamiento de los datos en cuatro grupos, ya que el quinto grupo tiene una cantidad de registros poco representativa. El resultado del coeficiente de Gini valida si existe una distribución adecuada de los datos y por lo que se observa al momento los datos no se encuentran distribuidos de manera normalizada.

Figura 14

Código del análisis estadístico, función de Gini.

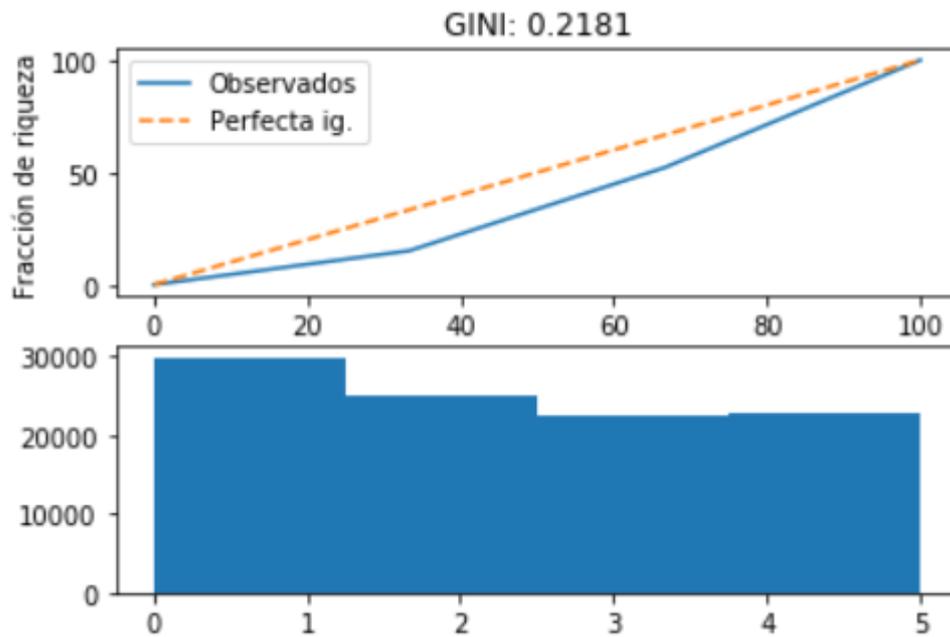
```
def G(v, n_bins):  
    #Añadirle variables para el el intervalo (min y max) y luego los n_bins.  
    bins = np.linspace(0, 100, n_bins)  
    total = float(np.sum(v))  
    yvals = []  
    for b in bins:  
        bin_vals = v[v <= np.percentile(v, b)]  
        bin_fraction = (np.sum(bin_vals) / total) * 100.0  
        yvals.append(bin_fraction)  
    # perfect equality area  
    pe_area = np.trapz(bins, x=bins)  
    # Lorenz area  
    lorenz_area = np.trapz(yvals, x=bins)  
    gini_val = (pe_area - lorenz_area) / float(pe_area)  
    return bins, yvals, gini_val
```

Nota. Elaboración del autor.

A continuación, se muestra de manera gráfica (figura 15) el coeficiente de Gini, luego del extenso análisis estadístico previo, se pudo determinar que para esta oferta comercial de paquetes prepago la cantidad de contenedores que demuestra un bajo coeficiente de Gini, es decir, es la mejor condición para evitar una distribución inequívoca de los ingresos, lo cual permite interpretar que con la cantidad de contenedores (4) utilizados como parámetro logra una distribución normalizada de los datos.

Figura 15

Gráfico del análisis estadístico, curva de Lorenz y coeficiente de Gini.



Nota. Elaboración del autor.

Luego del análisis, se generan varios insights; se puede obtener las generaciones de los usuarios a partir de la fecha de nacimiento una variable que agrega valor al sistema de recomendación. También se genera la variable `clt_ingresos` que permite diferenciar los 5 grupos para luego agregar los datos que permitan tener una distribución normalizada del dataset.

Figura 16

Código de la transformación de los datos. Parte 1

```
Transformacion de los datos

[19]: df.drop(columns='Unnamed: 0',inplace=True)

[20]: df2.drop(columns='Unnamed: 0',inplace=True)

[21]: #Transformacion de La data de df (dataset con informacion demografica del clte)
df.drop_duplicates(subset="ID", keep="last", inplace=True)
df['FECHA_NAC'] = pd.to_datetime(df['FECHA_NAC'])
df['GENERACION'] = df.apply(clasificador_generacion, axis=1)
df['CLT_INGRESOS'] = df.apply(clasificador_ingresos, axis=1)
df2.drop(columns='PCK_TIPO', inplace=True)

#Subset para Los usuarios de input - Se evaluan Los clientes que hayan comprado 'PAQUETES $1' y tenga ingreso Maximos de $2.
df_subset = df[(df['MAX_PAQ']==1) & (df['INGRESO_MES']<=2)]

df.drop(columns=['FECHA', 'FECHA_NAC', 'ID_IDENTIFICACION', 'USER_ID', 'PROFESION', 'ESTADO_CIVIL', 'PROVINCIA_WORK',
                'CANTON_WORK', 'SITIO_WORK', 'SITIO_HOME', 'ID_NSE_WORK', 'MAX_PAQ', 'SUM_PAQ', 'CANT_PAQ', 'INGRESO_MES'], inplace=True)
df_subset.drop(columns=['FECHA', 'FECHA_NAC', 'ID_IDENTIFICACION', 'USER_ID', 'PROFESION', 'ESTADO_CIVIL', 'PROVINCIA_WORK',
                       'CANTON_WORK', 'SITIO_WORK', 'SITIO_HOME', 'ID_NSE_WORK', 'MAX_PAQ', 'SUM_PAQ', 'CANT_PAQ', 'INGRESO_MES'], inplace=True)
```

Nota. Elaboración del autor.

Figura 17

Código de la transformación de los datos. Parte 2

```
#Transformacion de La data de df (dataset con informacion transaccional del clte)
time_limit_m1 = datetime.datetime(2020, 1, 1, 0, 0) # Mes1 2020
time_limit_m2 = datetime.datetime(2020, 2, 1, 0, 0) # Mes2 2020
df2['FECHA_ACTIVACION'] = pd.to_datetime(df2['FECHA_ACTIVACION'])

mask = (df2['FECHA_ACTIVACION']>=time_limit_m1) & (df2['FECHA_ACTIVACION']<time_limit_m2)
time_limit_m1 = df2[mask]
time_limit_m1.drop(columns=['TIPO_PAQUETE'], inplace=True)

df_trans_mes_inicial = df2[df2['FECHA_ACTIVACION']<time_limit_m2]
df_trans_mes_2 = df2[df2['FECHA_ACTIVACION']>=time_limit_m2]
df_trans_mes_2.drop(columns=['TIPO_PAQUETE'], inplace=True)

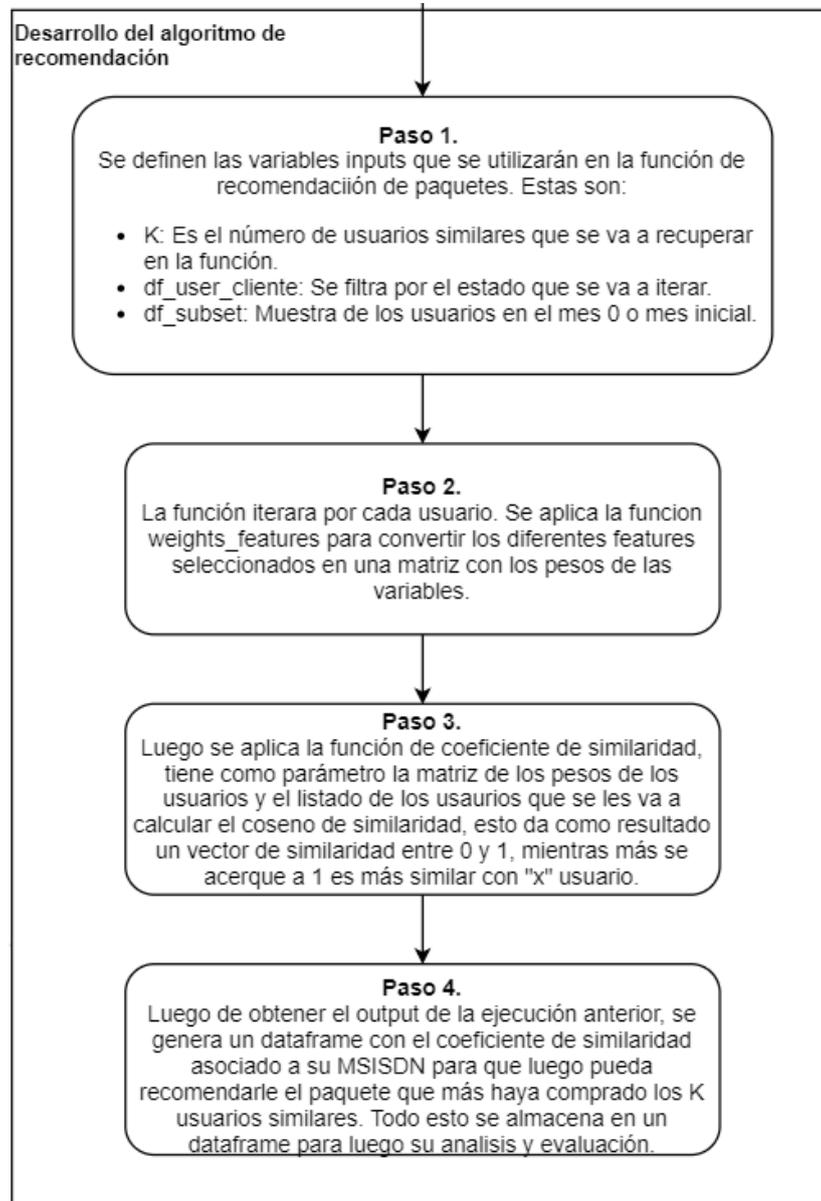
#El objetivo es evaluar a Los clientes del grupo PAQUETES $1 en Los grupos de paquetes de PAQUETES $3 Y PAQUETES $2.
#Filtro de La muestra para solo obtener grupo de usuarios con mayor simillaridad de Los grupos de paquetes de $2 y $3.
df_trans_mes_inicial = df_trans_mes_inicial[(df_trans_mes_inicial['GRUPO'] == 'PAQUETES $2') | (df_trans_mes_inicial['GRUPO'] == 'PAQUETES $3')]
```

Nota. Elaboración del autor.

Es importante particionar el dataset en varios meses, como se observa en la figura 16 y 17, para poder determinar el porcentaje de acierto de la recomendación a lo largo del tiempo. Es decir, se realiza un corte de un periodo de tiempo de 30 días, el cual se establece como el mes inicial y luego se realiza otro corte del siguiente mes para poder evaluar si la recomendación acierta.

Figura 18

Procesos del algoritmo de recomendación



Nota. Elaboración del autor.

Como se observa en la gráfica del flujo del proceso, figura 18, se requiere listar todos los features o características, las cuales definimos en la figura 19, que se van a utilizar para la elaboración de una matriz de similaridad, fundamental para elaborar la función de recomendación de paquetes. Es importante mencionar que la cantidad de features a utilizar dependerá del

criterio de negocio y a la disponibilidad de estos según el dataset. En este caso son limitados debido a que es un dataset con fines educacionales. Sin embargo, una empresa de telecomunicaciones debe tener una cantidad superior de features a los presentados en este trabajo.

**Figura 19**

*Código de la transformación de los datos. Parte 2*

```
# Lista de features que se utilizarán en la matriz de similaridad.
features = ['EDAD', 'PROVINCIA', 'NSE',
            'GENERO', 'CLT_INGRESOS', 'GENERACION']

# Lista de features categóricos que se utilizarán en la matriz de similaridad.
# Estas variables se encuentran también en la lista "features".

categorical_features = ['EDAD', 'PROVINCIA', 'NSE',
                        'GENERO', 'CLT_INGRESOS', 'GENERACION']
```

Nota. Elaboración del autor.

Por lo costoso del procesamiento se define las variables (figura 20) que se utilizarán como parámetro para la función de recomendación.

K: Cantidad de usuarios similares que va a determinar el algoritmo, sirve como base para luego recomendar el paquete más similar.

arr\_sub: arreglo que contiene un listado de los IDs que corresponden a los usuarios de la muestra estado que se va a recomendar.

**Figura 20**

*Variables de ingreso para la función de recomendación.*

```
#Para el testeo se puede delimitar la cantidad de usuarios a evaluar. Ya que el costo de procesamiento es alto.

K = 5 # Numero de usuario similares
cod_estado = '1'
arr_sub = df_subset[df_subset['ESTADO'] == cod_estado]
arr_sub_id = arr_sub['ESTADO']
df_usr_cliente = df.copy()
df_usr_cliente = df[df['ESTADO'] == cod_estado]
df_transaccionalidad = df_trans_mes_inicial.copy()
```

Nota. Elaboración del autor.

Se procede con la ejecución de la función de recomendación la cual se encarga de evaluar los usuarios input (en este caso usuarios de un estado en específico) contra la base de usuarios, iterará por cada usuario, de todos los estados y de esta manera obtener los usuarios más similares, de los cuáles se obtendrá el paquete que se va a recomendar, la cual permite mostrar cuánto tiempo se demorará y cuántos usuarios procesará y le otorgará una recomendación.

**Figura 21**

*Definición de la función de recomendación.*

```
res = recom_paquete_v2(df_subset, df_usr_cliente, df_transaccionalidad, features, categorical_features, K, provincia)
La funcion iterara: 123 veces.
100%|██████████| 123/123 [00:08<00:00, 14.66it/s]
fin de la funcion
```

Nota. Elaboración del autor.

Para poder evaluar el porcentaje de acierto del sistema de recomendación se compara el monto recomendado respecto a las transacciones realizadas por este usuario el siguiente mes y se le otorga el valor de 1 si la recomendación fue acertada y 0 si no acertó. En la figura 22 se puede interpretar que el sistema de recomendación tuvo un acierto aproximado del 87%, con un error del 13%. En caso de no ser satisfactorio el resultado para el negocio se vuelve a iterar (vuelve a comenzar desde 0 el algoritmo) en base a nuevos criterios.

**Figura 22**

*Definición de la función de recomendación*

```
df_trans_m2['RECOMENDACION'] = df_trans_m2.apply(clasificador_prediccion_paquete, axis=1)

df_trans_m2.head(10)
```

	MSISDN	FECHA_ACTIVACION	PCK_TIPO	MONTO	GRUPO	PAQUETE_RECOMENDADO	MONTO_RECOMENDADO	RECOMENDACION
0	47	2020-02-10 22:40:31	SVA_HUAWEI	3.0	PAQUETES \$3	PAQUETES \$2	2.0	0
1	47	2020-02-16 08:32:16	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
2	47	2020-02-28 20:38:22	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
3	47	2020-02-10 12:27:21	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
4	211	2020-02-23 11:11:26	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
5	211	2020-02-09 19:06:07	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
6	238	2020-02-29 09:10:52	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
7	238	2020-02-29 09:10:52	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
8	238	2020-02-09 21:33:15	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
9	238	2020-02-09 21:33:15	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1

```
df_trans_m2['RECOMENDACION'].value_counts(normalize=True)

1    0.870256
0    0.129744
Name: RECOMENDACION, dtype: float64
```

Resultado para la provincia 5, 87% de acierto.

Nota. Elaboración del autor.

Se realizan iteraciones adicionales en donde el sistema de recomendación arroja los siguientes resultados:

- Para el Anexo 3: el porcentaje de acierto es del 90% y el error es del 10%.
- Para el Anexo 4: el porcentaje de acierto es del 87% y el error es del 13%.
- Para el Anexo 5: la métrica de acierto es del 66.66% y el error es del 33.33%.

Si se calcula un promedio de las tres iteraciones realizadas se obtiene una métrica de acierto del 81%, siendo el 60% el mínimo viable para un sistema de recomendación, por lo que se considera el sistema exitoso.

## Conclusión

Para el presente trabajo de titulación se presentan las siguientes conclusiones obtenidas:

Se establece y elabora un flujograma con todos los procesos involucrados en el sistema de recomendación de paquetes, lo que permite identificar las variables esenciales para el sistema.

Se realiza un análisis estadístico basado en la curva de Lorenz y el coeficiente Gini, donde se pudo observar que el coeficiente más bajo es de 0.26 cuando se agrupa en cuatro (4) bins, lo que permite concluir que los datos se encuentran en un nivel de distribución normalizada uniforme aceptable para el sistema ya que si se agrupa de otra manera o no se realiza ninguna agrupación el coeficiente incrementa por lo que se tendrán recomendaciones con baja precisión.

Se considera factible, ya que el uso de la métrica coseno de similaridad permite obtener un acierto mayor al 80% para el sistema de recomendación basado en collaborative filtering.

Se diseñó un sistema de recomendación de paquetes prepago con enfoque de collaborative filtering donde el porcentaje de precisión promedio (basado en 3 iteraciones) es del 81% con un margen de error del 19%, lo que permite concluir que es un sistema eficaz al momento de recomendar paquetes a los usuarios.

## **Recomendaciones**

Para el presente trabajo de titulación se presentan las siguientes recomendaciones para las próximas líneas de investigación de este tema:

Se recomienda incrementar el uso de data histórica de los usuarios, además del uso de nuevas variables sociodemográficas para el sistema de recomendación ya que incrementará el nivel de similaridad de los usuarios.

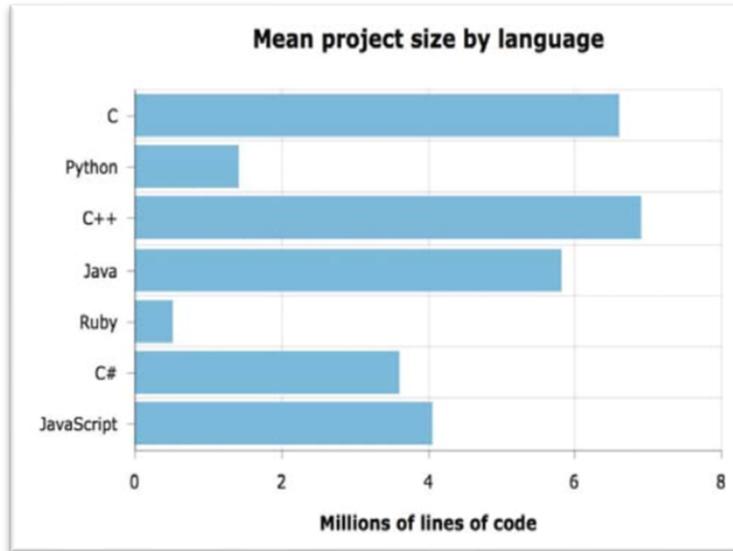
Se recomienda el diseño e implementación de sistemas recomendación de paquetes basados en collaborative filtering en las empresas de telecomunicaciones del Ecuador para ofrecer un servicio personalizado al usuario y mejorar su ARPU.

Se recomienda el diseño e implementación de sistemas recomendación de paquetes basados en collaborative filtering en otros modelos de negocio que cuenten con una base de datos enriquecida de sus usuarios.

Se recomienda la mezcla de varios enfoques de los sistemas de recomendación para mejorar la precisión del sistema en nuevas líneas de investigación.

# ANEXOS

## Anexo 1



Anexo 1 Tamaño medio por Proyecto. Tomado de (Nagpal & Gabrani, 2019)

## Anexo 2

Language	Time	Speedup
Python	6.39 ns	1x
Compiled Python	4.1 ns	1.56x
Cython	167 ns	38.26

Anexo 2 Comparación de velocidad entre Python y Cython para encontrar la serie Fibonacci. Tomado de (Nagpal & Gabrani, 2019)

## Anexo 3

```
df_trans_m2['RECOMENDACION'] = df_trans_m2.apply(clasificador_prediccion_paquete, axis=1)
```

```
df_trans_m2.head(10)
```

	MSISDN	FECHA_ACTIVACION	PCK_TIPO	MONTO	GRUPO	PAQUETE_RECOMENDADO	MONTO_RECOMENDADO	RECOMENDACION
0	371	2020-02-04 11:32:21	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
1	371	2020-02-09 18:17:43	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
2	371	2020-02-29 17:34:48	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
3	465	2020-02-05 14:28:41	GPRS	2.0	PAQUETES \$2	PAQUETES \$3	3.0	1
4	465	2020-02-05 14:28:41	GPRS	2.0	PAQUETES \$2	PAQUETES \$3	3.0	1
5	465	2020-02-05 14:28:41	GPRS	2.0	PAQUETES \$2	PAQUETES \$3	3.0	1
6	465	2020-02-05 14:28:41	GPRS	2.0	PAQUETES \$2	PAQUETES \$3	3.0	1
7	465	2020-02-10 15:58:50	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
8	465	2020-02-10 15:58:50	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
9	465	2020-02-16 13:29:30	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1

```
df_trans_m2['RECOMENDACION'].value_counts(normalize=True)
```

```
1    0.9
0    0.1
Name: RECOMENDACION, dtype: float64
```

Resultado para la provincia 6, 90% de acierto.

*Anexo 3 Recomendación en base a la provincia 6. Elaboración propia*

## Anexo 4

```
df_trans_m2['RECOMENDACION'] = df_trans_m2.apply(clasificador_prediccion_paquete, axis=1)
```

```
df_trans_m2.head(10)
```

	MSISDN	FECHA_ACTIVACION	PCK_TIPO	MONTO	GRUPO	PAQUETE_RECOMENDADO	MONTO_RECOMENDADO	RECOMENDACION
0	47	2020-02-10 22:40:31	SVA_HUAWEI	3.0	PAQUETES \$3	PAQUETES \$2	2.0	0
1	47	2020-02-16 08:32:16	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
2	47	2020-02-28 20:38:22	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
3	47	2020-02-10 12:27:21	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
4	211	2020-02-23 11:11:26	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
5	211	2020-02-09 19:06:07	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
6	238	2020-02-29 09:10:52	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
7	238	2020-02-29 09:10:52	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
8	238	2020-02-09 21:33:15	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
9	238	2020-02-09 21:33:15	DATOS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1

```
df_trans_m2['RECOMENDACION'].value_counts(normalize=True)
```

```
1    0.870256
0    0.129744
Name: RECOMENDACION, dtype: float64
```

Resultado para la provincia 5, 87% de acierto.

*Anexo 4 Recomendación en base a la provincia 5. Elaboración propia*

## Anexo 5

```
df_trans_m2['RECOMENDACION'] = df_trans_m2.apply(clasificador_prediccion_paquete, axis=1)
df_trans_m2.head(10)
```

	MSISDN	FECHA_ACTIVACION	PCK_TIPO	MONTO	GRUPO	PAQUETE_RECOMENDADO	MONTO_RECOMENDADO	RECOMENDACION
0	4016	2020-02-09 22:41:04	GPRS	2.0	PAQUETES \$2	PAQUETES \$2	2.0	0
1	4016	2020-02-01 22:08:59	GPRS	1.0	PAQUETES \$1	PAQUETES \$2	2.0	1
2	11088	2020-02-04 10:11:18	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
3	23612	2020-02-05 18:05:22	DATOS	2.0	PAQUETES \$2	PAQUETES \$3	3.0	1
4	23612	2020-02-23 19:32:08	GPRS	3.0	PAQUETES \$3	PAQUETES \$3	3.0	0
5	23612	2020-02-01 13:58:33	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
6	23612	2020-02-29 20:50:53	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
7	23612	2020-02-04 08:03:34	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1
8	24389	2020-02-28 11:15:39	GPRS	3.0	PAQUETES \$3	PAQUETES \$3	3.0	0
9	27497	2020-02-20 18:28:12	GPRS	1.0	PAQUETES \$1	PAQUETES \$3	3.0	1

```
df_trans_m2['RECOMENDACION'].value_counts(normalize=True)
```

```
1    0.666667
0    0.333333
Name: RECOMENDACION, dtype: float64
```

Resultado para la provincia 22, 66.66% de acierto.

*Anexo 5 Recomendación en base a la provincia 22. Elaboración propia*

## REFERENCIAS

- Alves, D., Valente, B., Filipe, R., Castro, M., & Macedo, L. (30 de Agosto de 2019). *Springer Link*. Obtenido de A Recommender System for Telecommunication Operators' Campaigns: [https://link.springer.com/chapter/10.1007/978-3-030-30244-3\\_8#citeas](https://link.springer.com/chapter/10.1007/978-3-030-30244-3_8#citeas)
- Balmer, R., Stanford, L., & Schmindt, S. (Julio de 2020). *ScienceDirect*. Obtenido de Artificial Intelligence Applications in Telecommunications and other network industries: <https://www.sciencedirect.com/science/article/abs/pii/S0308596120300690>
- Golavatchev, J., Buddle, O., & Hong. (1 de January de 2010). *ResearchGate*. Obtenido de Next Generation Telco Product Lifecycle Management - How to Overcome Complexity in Product Management by Implementing Best-Practice PLM: [https://www.researchgate.net/publication/204100092\\_Next\\_Generation\\_Telco\\_Product\\_Lifecycle\\_Management\\_-\\_How\\_to\\_Overcome\\_Complexity\\_in\\_Product\\_Management\\_by\\_Implementing\\_Best-Practice\\_PLM](https://www.researchgate.net/publication/204100092_Next_Generation_Telco_Product_Lifecycle_Management_-_How_to_Overcome_Complexity_in_Product_Management_by_Implementing_Best-Practice_PLM)
- GSMA. (13 de Diciembre de 2019). *GSMA*. Obtenido de The Mobile Economy Latin America 2019: <https://www.gsma.com/latinamerica/resources/the-mobile-economy-latin-america-2019/>
- Han, J., Li, Z., Martin, K., Scott, H., David, G. S., Zaria, T., . . . J. Feuer, E. (3 de Agosto de 2016). *Springer*. Obtenido de Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics: <https://link.springer.com/content/pdf/10.1186/s12942-016-0056-6.pdf>
- Hernández, R., Fernández, C., & Del Pilar, M. (Abril de 2014). *Universiddad florencio del castillo*. Obtenido de <https://www.uca.ac.cr/wp-content/uploads/2017/10/Investigacion.pdf>

- Jansen, P. (Julio de 2020). *TIOBE*. Obtenido de Tioebe the software quality company:  
<https://www.tiobe.com/tiobe-index/>
- Jiayin Qi, F. W. (17 de Agosto de 2007). *Wiley Online Library*. Obtenido de Artificial intelligence applications in the telecommunications industry:  
<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0394.2007.00433.x>
- Koren, Y., Bell, R., & Volinsky, C. (07 de Agosto de 2009). *IEEE*. Obtenido de Matrix Factorization Techniques for Recommender Systems:  
<https://ieeexplore.ieee.org/abstract/document/5197422/>
- N. Kok, J., Egbert, J. W., A. Koster, W., & van der Putten, P. (2009). *ARTIFICIAL INTELLIGENCE: DEFINITION, TRENDS*. Obtenido de Artificial Intelligence : Definition, Trends, Techniques and Cases.
- Nagpal, A., & Gabrani, G. (29 de Abril de 2019). *IEEE*. Obtenido de Python for Data Analytics, Scientific and Technical Applications:  
<https://ieeexplore.ieee.org/abstract/document/8701341>
- O'Dea, S. (27 de Febrero de 2020). *Statista*. Obtenido de Mobile data usage per user per month worldwide 2018-2024, by region:  
<https://www.statista.com/statistics/489169/canada-united-states-average-data-usage-user-per-month/>
- Orozco, M. (2 de Marzo de 2019). *El Comercio*. Obtenido de Lluvia de ofertas con internet móvil para impulsar ventas: <https://www.elcomercio.com/actualidad/planes-ofertas-internet-movil-telefonos.html>
- Piatetsky, G. (Diciembre de 2018). *KDnuggets*. Obtenido de Here are the most popular Python IDEs / Editors: <https://www.kdnuggets.com/2018/12/most-popular-python-ide-editor.html>

- Rossetti, M., Stella, F., & Zanker, M. (08 de Octubre de 2013). *IEEE Xplore*. Obtenido de Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems: <https://ieeexplore.ieee.org/abstract/document/6621365>
- Satari, A. (23 de Junio de 2020). *GSMA*. Obtenido de Region in Focus: Latin America, Q1 2020: <https://www.gsma.com/latinamerica/resources/region-in-focus-latin-america-q1-2020/>
- Sondur, S., Chigadani, A., & Nayak, S. (Mayo de 2016). *SemanticScholar*. Obtenido de Similarity Measures for Recommender Systems: A Comparative Study: <https://pdfs.semanticscholar.org/010c/f75eaae88fc7402f3ba62949e377d701190a.pdf>
- W. Murdoch, J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (14 de Enero de 2019). *Cornell University*. Obtenido de Interpretable machine learning: definitions, methods, and applications: <https://arxiv.org/abs/1901.04592>
- Wu, D., Zeng, Zeng, G., Meng, L., Zhou, W., & Li, L. (25 de Enero de 2017). *IEEE Xplore*. Obtenido de Gini Coefficient-based Task Allocation for Multi-robot Systems with Limited Energy Resources: <https://ieeexplore.ieee.org/abstract/document/7833258>



**Presidencia  
de la República  
del Ecuador**



**Plan Nacional  
de Ciencia, Tecnología,  
Innovación y Saberes**



**SENESCYT**  
Secretaría Nacional de Educación Superior,  
Ciencia, Tecnología e Innovación

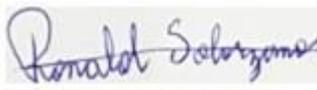
## **DECLARACIÓN Y AUTORIZACIÓN**

Yo, **Solórzano Navarrete Ronald David**, con C.C: # **0930408992** autor/a del trabajo de titulación: **Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador** previo a la obtención del título de **Ingeniero en Sistemas Computacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, **22 de septiembre de 2020**

f. 

**Ronald David Solórzano Navarrete**

**C.C: 0930408992**

## REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

### FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	Diseño de un sistema de recomendaciones de paquetes prepago mediante uso de técnicas de inteligencia artificial en redes telefónicas en Ecuador		
AUTOR(ES)	Ronald David Solórzano Navarrete		
REVISOR(ES)/TUTOR(ES)	Gustavo Molina Flores		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Ingeniería		
CARRERA:	Ingeniería en Sistemas computacionales		
TÍTULO OBTENIDO:	Ingeniero en Sistemas computacionales		
FECHA DE PUBLICACIÓN:	22 de septiembre de 2020	No. DE PÁGINAS:	81
ÁREAS TEMÁTICAS:	Sistema de recomendación, inteligencia artificial		
PALABRAS CLAVES/KEYWORDS:	Machine Learning, Sistema de Recomendación, métodos estadísticos, coseno de similaridad, Python, Collaborative Filtering.		
RESUMEN/ABSTRACT:	<p>En este proyecto de titulación trata sobre realizar con el objetivo desarrollar e implementar para el control del personal que realiza delivery o motorizado con el fin de mejorar u optimizar el proceso de trabajo que labora la empresa PROCO C. Ltda. y encontrar a solucionar los problemas ineficientes que tienen la empresa y poder dar un buen sistema eficiente de su proceso de trabajo. Para esto se necesita saber funcionamiento del trabajo que actualmente se hace, como el proceso de tomar el pedido a los clientes y el proceso del motorizado, al saber el proceso, se actúa a analizar los pequeños problemas que se encuentra en el proceso de trabajo que hacen en la empresa PROCO C. Ltda. Por lo tanto, al tomar los problemas que se presentan, ca a tomar como requisitos para el desarrollo de software y así crear un sistema mejor para la empresa.</p>		
ADJUNTO PDF:	SI	NO	
CONTACTO CON AUTOR/ES:	Teléfono: +593958881784	E-mail: <a href="mailto:ronald_dsn95@hotmail.com">ronald_dsn95@hotmail.com</a>	
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UTE)::	Nombre: EDISON JOSE TOALA QUIMI		
	Teléfono: +593990976776		
	E-mail: <a href="mailto:edison.toala@cu.ucsg.edu.ec">edison.toala@cu.ucsg.edu.ec</a>		
<b>SECCIÓN PARA USO DE BIBLIOTECA</b>			
Nº. DE REGISTRO (en base a datos):			
Nº. DE CLASIFICACIÓN:			
DIRECCIÓN URL (tesis en la web):			