

# **UNIVERSIDAD MAYOR DE SAN ANDRÉS**

**FACULTAD DE CIENCIAS ECONÓMICAS Y  
FINANCIERAS**

**CARRERA DE ADMINISTRACIÓN DE EMPRESAS  
INSTITUTO DE INVESTIGACIÓN Y CAPACITACIÓN EN  
CIENCIAS ADMINISTRATIVAS - IICCA**



## **INTELIGENCIA ARTIFICIAL Y SU INCIDENCIA EN LA IDENTIFICACIÓN DE PERFILES DE SUSCRIPCIÓN EN LA EDUCACIÓN VIRTUAL**

**Tesis de Postgrado para obtener el Título de Magíster Scientiarum**

**POR: RODRIGO HUGO AVALOS QUISPE**

**TUTOR: WILFREDO MATÍAS POMA**

**LA PAZ - BOLIVIA  
Septiembre, 2021**



## **Dedicatoria**

Esta tesis la dedico a mi madre que estuvo siempre a mi lado brindándome su mano amiga dándome a cada instante una palabra de aliento para llegar a cumplir las metas propuestas por mi persona, sus esfuerzos son el aliciente para seguir mejorando, mi inspiración Teodora Valeria Quispe Cruz, gracias.



### **Agradecimiento**

A las personas que me ayudaron a comprender estas herramientas para nuestra era actual, la era digital. Trabajar todos estos elementos son el fruto del esfuerzo y perseverancia, agradecerle al Msc. Wilfredo Matias por su tiempo y dedicación en la elaboración del presente trabajo.



## Índice

RESUMEN .....	2
CAPÍTULO I .....	3
ANTECEDENTES .....	3
1.1. Definición conceptual del área de investigación .....	3
1.2. Identificación y formulación del problema de investigación .....	4
1.2.1. Formulación del problema.....	5
1.3. OBJETIVOS DE LA INVESTIGACIÓN .....	5
1.3.1. Objetivo general .....	5
1.3.2. Objetivos específicos.....	5
1.4. Justificación de la investigación.....	6
1.4.1. Justificación teórica.....	6
1.4.2. Justificación metodológica .....	7
1.4.3. Justificación práctica.....	9
1.4.4. Justificación social .....	10
1.5. Alcances de la investigación .....	11
1.5.1. Ámbito geográfico .....	11
1.5.2. Ámbito político, social o económico .....	11
1.5.3. Ámbito industrial/sectorial .....	12
1.5.4. Ámbito financiero .....	12
CAPITULO II.....	14
MARCO TEÓRICO.....	14
2.1. Modelaje de inteligencia artificial .....	14
2.1.1. Aprendizaje automático (Machine Learning) .....	14
2.1.2. Aprendizaje supervisado .....	16
2.1.3. Aprendizaje no supervisado.....	16
2.1.4. Constatación del problema .....	17
2.1.5. Conjunto de datos de entrenamiento .....	19
2.2. MODELACIÓN DE LA FUNCIÓN OBJETIVO.....	20
2.2.1. Función de costos (minimización) .....	20
2.2.2. Función de mínimos cuadrados.....	21



2.2.3. Gradiente de descenso .....	23
2.2.4. Regresión logística .....	24
2.2.5. Modificación de parámetros $\Theta_0$ y $\Theta_1$ .....	25
2.2.6. Gradiente de descenso .....	26
2.3. PERFILES DE CONSUMIDOR.....	26
2.3.1. Información de perfiles potenciales.....	26
2.3.2. Actualización y asignación de perfiles .....	28
2.3.3. Recomendaciones al momento de asignar etiquetas al perfil del cliente actual .....	29
2.3.4. Toma de decisiones individual.....	29
2.3.5. Influencia externa.....	31
CAPITULO III.....	34
METODOLOGÍA DE LA INVESTIGACIÓN.....	34
3.1. Tipo de análisis .....	34
3.2. Marco metodológico.....	34
3.3. Diseño de la investigación.....	34
3.4. Unidades de análisis.....	34
3.5. Determinación del universo y tamaño de la muestra.....	35
3.6. Hipótesis de estudio.....	37
3.7. Indicadores de estudio.....	37
3.8. Operacionalización de variables.....	40
3.9. Unidades de análisis.....	41
CAPITULO IV .....	42
MARCO PRÁCTICO .....	42
4.1. Análisis de campo .....	42
4.1.1. Análisis de las variables .....	44
4.1.2. Constatación del problema en base a las variables validadas .....	48
4.2. Resultados.....	50
4.3. Comprobación de hipótesis.....	54
CAPITULO V.....	58
PROPUESTA.....	58



5.1. PROPUESTA – MODELADO Y EJECUCIÓN DE UN MODELO DE MACHINE LEARNING APLICADO A LA IDENTIFICACIÓN DE PERFILES DE SUSCRIPCIÓN EN LA EDUCACIÓN VIRTUAL EN UNA ORGANIZACIÓN CON FINES DE LUCRO .....	58
5.1.1. Justificación .....	58
5.1.2. Objetivo general .....	58
5.1.3. Propuesta de programa de identificación de suscriptores en una academia de educación virtual.....	60
5.2. PROCESO OPERATIVO 3 .....	61
5.2.1. Importando el DataFrame .....	61
Importando las siguientes librerías: .....	61
5.2.2. Visualización del conjunto de datos (dataset) .....	61
5.3. PROCESO OPERATIVO 4 .....	63
5.3.1. Declarar las variables de trabajo .....	63
5.3.2. Configurar el árbol de decisiones .....	63
5.3.3. Modelado .....	64
5.3.4. Predicción .....	64
5.3.5. Evaluación .....	65
5.3.6. Visualización.....	65
5.3.7. Interpretación.....	66
5.4. PROCESO ADMINISTRATIVO 5 .....	67
CAPITULO VI .....	69
CONCLUSIONES Y RECOMENDACIONES.....	69
6.1. Conclusiones .....	69
6.2. Recomendaciones .....	72
REFERENCIAS .....	74
ANEXOS .....	75



## Índice de tablas

Tabla 1. Construcción de indicadores de entrada .....	35
Tabla 2. Operacionalización de las variables de estudio .....	40
Tabla 3. Variables de modelado .....	48
Tabla 4. Matriz de confusión .....	49
Tabla 5. Nivel de conteo y predicción .....	49
Tabla 6. Resultados de entrenamiento .....	50
Tabla 7. Nivel de Gini, entropía y diferencia.....	52
Tabla 8. Importar librerías a emplear en Python .....	61
Tabla 9. Visualización de los datos correspondientes al archivo entrenamiento 2020 .....	62
Tabla 10. Descriptivos del dataframe .....	62
Tabla 11. Importando el algoritmo de árbol de decisión .....	63
Tabla 12. Visualización de los elementos y variables en el modelo de árboles de decisión .....	64
Tabla 13. Modelado del árbol de decisiones .....	64
Tabla 14. Determinación de los niveles y parámetros en la ejecución del modelo ..	65
Tabla 15. Evaluación del modelo .....	65
Tabla 16. Visualización del árbol de decisión construido empleando el lenguaje de programación Python.....	66



## Índice de figuras

Figura 1. Representación de una neurona artificial simple de 3 capas.....	7
Figura 2. El nuevo paradigma de programación: <i>Aprendizaje automático</i> .....	15
Figura 3. Parámetros de análisis de conjunto de datos.....	17
Figura 4. Visualización de (IA) matriz de confusión.....	20
Figura 5. Visualización de la función de costos.....	22
Figura 6. Ilustración del mínimo local.....	24
Figura 7. Visualización del gradiente de descenso .....	26
Figura 8. Agrupación de perfiles de suscripción.....	28
Figura 9. Representación del proceso de compra del consumidor .....	30
Figura 10. Clasificación de productos según los niveles de abstracción .....	31
Figura 11. Árbol de decisión en base a la clasificación de etiquetas .....	43
Figura 12. Datos categóricos - género .....	44
Figura 13. Porcentajes correspondientes - género.....	44
Figura 14. Datos categóricos - nivel educativo.....	45
Figura 15. Porcentajes correspondientes – nivel educativo .....	45
Figura 16. Frecuencia - etiqueta edad .....	46
Figura 17. Distribución - etiqueta ciudad.....	47
Figure 18. Probabilidad de distribución .....	47
Figura 19. Nivel de errores .....	50
Figura 20. Nivel de aprendizaje del algoritmo .....	51
Figura 21. Visualización del árbol de decisión aplicado .....	54
Figura 22. Resumen de la validación del algoritmo de inteligencia artificial de árboles de decisión .....	55
Figura 23. Situación de la asignación de recursos publicitarios y el ROI antes de la adopción de un algoritmo de inteligencia artificial para la identificación de perfiles en la educación virtual.....	56
Figura 24. Impactos de la implantación de un algoritmo basado en árboles de decisión en la creación de estrategias de asignación de recursos en el marketing digital.....	57
Figura 25. Propuesta de adopción del aprendizaje automático en una organización con fines de lucro.....	59





Figura 26. Propuesta de programa de identificación de suscriptores en la educación virtual (código fuente para el lenguaje de programación Python) .....	60
Figura 27. Proceso administrativo 5 .....	67

**“Inteligencia artificial y su incidencia en la identificación de  
perfiles de suscripción en la educación virtual”**



## RESUMEN

El presente trabajo se enfoca la utilización de un modelo de inteligencia artificial aplicada a la educación virtual para crear un factor diferenciador en la toma de decisiones, la asignación de recursos, el control de la calidad servicios, entre otros, haciéndola más eficiente, en un mundo cada vez más dinámico, de continuos y rápidos cambios y cada vez más competitivo. El uso de las tecnologías de la información y las comunicaciones como producto de la inteligencia artificial se ha convertido también en herramienta fundamental para el éxito de las organizaciones. En el primer capítulo se presenta los objetivos de la presente investigación con el propósito de definir el campo de acción basado en que la “inteligencia artificial y su incidencia en la identificación de perfiles de suscripción en la educación virtual”, posteriormente se hace una breve conceptualización de los argumentos teóricos que están enfocados a describir los principios de inteligencia artificial y los algoritmos de clasificación en este caso se adoptó un modelo de clasificación basado en arboles de decisión; que nos hará posible entender el estado del arte, en el tercer capítulo se enmarca a definir las variables de trabajo y sus desencadenantes, para poder encontrar las inferencias y descriptivos correspondientes al capítulo cuarto, y concluir con una propuesta de basada en los lineamientos de la gestión de la adopción de estos instrumentos como es la de la inteligencia artificial basada en un algoritmo de clasificación y aplicarlo en una organización que utiliza un sitio web como plataforma de posicionamiento.



## CAPÍTULO I

### ANTECEDENTES

#### 1.1. Definición conceptual del área de investigación

En el actual mundo globalizado, en donde se busca la optimización del uso de los recursos disponibles y competir no solo a nivel nacional sino también a nivel internacional impulsados por el potencial uso y desarrollo de la web, se vuelve una prioridad que las empresas trabajen en un nivel de gestión de alto nivel ya en el entendido de que existen modelos matemáticos que coadyuvan con esta tarea; en los diferentes niveles estratégico, táctico y operativo, con el fin de alcanzar estándares de eficiencia en la asignación de recursos, y poder optimizar la rentabilidad y/o beneficios de la empresa o unidad productiva.

Muchos de los estrategias de marketing o medios de organizaciones de cualquier nivel de Sud América, están adoptando el marketing digital dentro de sus planes generales. Sin embargo, en muchas ocasiones se hace uso de este solo como una herramienta más dentro de la estrategia de comunicación, desconociendo el alcance que se puede obtener si se desarrolla de forma estructurada con un modelo matemático cada uno de ellos con un parámetro de medición, no solo de resultados sino de generar un mayor tráfico en un sitio web específico.

Determinar una estrategia de marketing y objetivos, desde una perspectiva estratégica, para responder a las necesidades de una organización que persigue fines de lucro, puede ocasionar bifurcaciones tanto a nivel táctico y operativo. Un problema relacionado es que, en la medida que aumentan la complejidad y la especialización del comercio electrónico, es más difícil asignar los recursos disponibles a las diferentes actividades relacionadas al marketing de la manera más eficaz para la organización como un todo. Este



tipo de problemas y la necesidad de encontrar la mejor forma de resolverlos crearon el ambiente propicio para el surgimiento de la inteligencia artificial.

El presente trabajo se enfoca la utilización de un software de inteligencia artificial aplicada a la educación online para crear un factor diferenciador para la toma de decisiones, la asignación de recursos, el control de la calidad servicios, entre otros, haciéndola más eficiente, en un mundo cada vez más dinámico, de continuos y rápidos cambios y cada vez más competitivo. El uso de las tecnologías de la información y las comunicaciones como producto de la inteligencia artificial se ha convertido también en herramienta fundamental para el éxito de las organizaciones.

## **1.2. Identificación y formulación del problema de investigación**

El entorno que tiene un impacto representativo en la actualidad, además de tener una presencia cotidiana con mayor frecuencia; viene siendo el comercio digital. El cual se compone de diversos componentes, pero en especial los sitios web que representan un medio o herramienta básica para el desarrollo de un comercio electrónico.

La labor ha consistido en investigación-formación-desarrollo de un software basada en algoritmos heurísticos que encontrará los perfiles idóneos de potenciales suscriptores. Donde esta herramienta está enfocada a aumentar el número de suscriptores en la academia online Rodrigo Avalos (<https://rodrigo-avalos.com/>). Que a su vez genera información mediante indicadores que nosotros programamos y/o necesitamos para tomar decisiones. Debido a que este mecanismo tiene mayor importancia e incidencia en el tráfico web, la presente investigación introducirá un software de inteligencia artificial para la sistematización y posterior toma de decisiones para la optimización de asignación de recursos en campañas digitales que incide en el incremento del tráfico web, en consecuencia, los sitios web de



educación serán los ejes que otorgaran información primaria para la consecución de los objetivos propuestos en la presente investigación.

### **1.2.1. Formulación del problema**

La formulación del problema se origina en:

**¿Qué algoritmo de machine learning incide en la identificación de perfiles de suscripción en la educación virtual?**

## **1.3. OBJETIVOS DE LA INVESTIGACIÓN**

### **1.3.1. Objetivo general**

- Establecer un modelo de machine learning basado en la estructura de árboles de decisión para la identificación de perfiles de suscripción en la educación virtual.

### **1.3.2. Objetivos específicos**

- Definir la estructura del conjunto de entrenamiento a aplicar en un algoritmo de aprendizaje según los requerimientos de la educación virtual.
- Validar el algoritmo de aprendizaje con los parámetros requeridos por la educación virtual.
- Aplicar el algoritmo de aprendizaje en una plataforma de enseñanza online.
- Diseñar un programa de identificación de suscriptores en una academia de educación virtual.



## **1.4. Justificación de la investigación**

### **1.4.1. Justificación teórica**

La inteligencia artificial (IA) tiene por objetivo el estudio y el análisis del comportamiento humano en los ámbitos de la comprensión, de la percepción, de la resolución de problemas y de la toma de decisiones con el fin de poder reproducirlos con la ayuda de un computador. De esta manera, las aplicaciones de la (IA) se sitúan principalmente en la simulación de actividades intelectuales del hombre (Turing, 1954). Es decir, imitar por medio de máquinas, normalmente electrónicas, tantas actividades mentales como sea posible, y quizás llegar a mejorar las capacidades humanas en estos aspectos y generar mayor eficiencia en todo el proceso que supone la aplicación de la inteligencia artificial en sí.

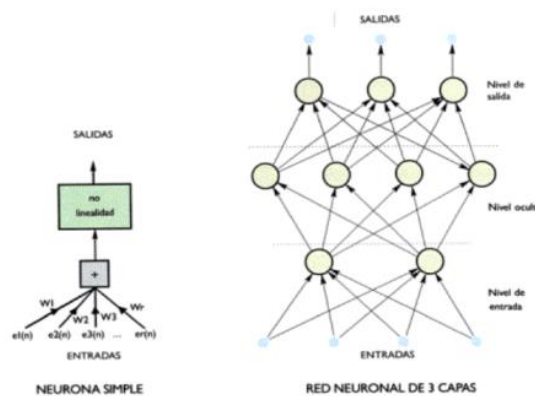
Por otro lado, según (Bourcier, 2003, p.56) dice que la inteligencia artificial es una rama de la informática que intenta reproducir las funciones cognitivas humanas como el razonamiento, la memoria, el juicio o la decisión y, después, confiar una parte de esas facultades, que se consideramos signos de inteligencia, a los ordenadores.

Así entonces, se puede afirmar que la Inteligencia Artificial es una disciplina eminentemente tecnológica que persigue la construcción de máquinas y programas capaces de realizar complejas tareas con una habilidad y eficiencia iguales o superiores a las que consigue el ser humano. La Inteligencia Artificial definida por John Mc Carthy, (1956) como “la ciencia y la ingeniería de hacer inteligentes a las máquinas” aplicada a los negocios en la organización puede brindar grandes ventajas competitivas, mejorar los procesos al interior de la organización e incluso reducir costos en las operaciones y servir de base para la toma de decisiones adecuadas. Existen muchas características que hacen que esta herramienta pueda mostrar excelentes condiciones de ser una tecnología aplicable, que coadyuvara a que las empresas sean más productivas. Según el análisis, los datos, aplicación de un algoritmo y

detección de comportamientos latentes, estos procesos tienen una importancia significativa para la toma de decisiones tanto tácticas y estratégicas, empero, en este caso se da la importancia de a la asignación de recursos en campañas digitales.

En conclusión, las expectativas de la Inteligencia Artificial, están basadas en el hecho de verla como una opción de inversión en el desarrollo de herramientas que permitan reducir costos, distribuir recursos, detección de fraudes, ayuda en línea, captura de conocimiento, etc. Hernández R. (2005). En base a que las redes neuronales trabajan como filtros de tamización y clasificación de los datos, según los datos con los que se alimente el sistema, siendo así su mayor o menor complejidad de aprendizaje.

**Figura 1.** Representación de una neurona artificial simple de 3 capas



**Fuente:** Proceso de transmisión de la información en una red neuronal.

### 1.4.2. Justificación metodológica

Así mismo, la Inteligencia Artificial aplicada a sistemas de control de calidad en las empresas, ha permitido minimizar e incluso evitar las devoluciones de productos por defectos o problemas de calidad, debido a que detectan con mayor precisión los productos no conformes, mejorando la percepción de calidad en los clientes, creando recordación de marca y posicionando mejor a





las compañías. Además, estos sistemas permiten ahorro de tiempo, exactitud, reducción de costos puesto que la mano de obra es más costosa en términos de carga prestacional y es más susceptible a equivocaciones. Osuna, S. (2008).

La Inteligencia Artificial es fundamental como apoyo en la toma de decisiones en las organizaciones, especialmente en el área financiera y administrativa, ya que coadyuva proveyendo bases sólidas, fiables y seguras para tomar decisiones adecuadas y acertadas para el cumplimiento de los objetivos corporativos.

Desde un punto de vista ideal, los objetivos formulados deben coincidir con los de toda la organización; sin embargo, esta coincidencia no siempre es conveniente. Muchos problemas interesan sólo a una parte de la organización, de manera que el análisis sería demasiado extenso si los objetivos fueran generales y se prestara atención especial a todos los efectos secundarios sobre el resto de la organización. En lugar de ello, los objetivos de un estudio deben ser tan específicos como sea posible, siempre y cuando consideren las metas principales del tomador de decisiones y mantengan un nivel razonable de congruencia con los objetivos de niveles más elevados.

Donde la principal causa toma forma, en un objetivo estratégico de incrementar las suscripciones de una academia online, que a su vez son parámetros de gestión propia de un prestador de servicios educativos académicos y como así también la asignación de recursos a la captación de los mismos. Es fundamental encontrar herramientas que contribuyan a una asignación de recursos económicos en base a el análisis predictivo y basados en algoritmos acordes a las características de los datos generados en base al tráfico en una academia de educación virtual.



### 1.4.3. Justificación práctica

Como herramienta de toma de decisiones, la (IA) es tanto una ciencia como un arte. Es una ciencia por las técnicas matemáticas que incorpora, y un arte porque el éxito de las fases que conducen a la solución del modelo depende en gran medida de la creatividad y experiencia del equipo de IA. Willemain (1994) manifiesta que “una práctica de (IA) eficaz requiere más que competencia analítica. También requiere, entre otros atributos, juicio técnico (es decir, cuándo y cómo utilizar una técnica dada), así como habilidades de comunicación y supervivencia organizacional”.

La inteligencia artificial ofrece un gran valor añadido que puede transformar una organización desde la óptica estratégica y operativa que hace que la implementación de un algoritmo de (IA) pueda generar tanto un aprendizaje como organización como el de crear una herramienta de gestión de datos. Sin embargo, en la práctica, muchas organizaciones lucrativas no utilizan este enfoque. Algunos estudios de corporaciones estadounidenses han demostrado que la administración tiende a adoptar la meta de ganancias satisfactorias combinada con otros objetivos, en lugar de enfocarse en la maximización de la ganancia a largo plazo. Algunos de estos otros objetivos pueden ser conservar la estabilidad de las ganancias, aumentar o conservar la participación de mercado con que se cuenta, permitir la diversificación de productos, mantener precios estables, mejorar las condiciones y el ánimo de los trabajadores, mantener el control familiar sobre el negocio o incrementar el prestigio de la compañía. Si se satisfacen estos objetivos, tal vez se logre maximizar las ganancias a largo plazo, pero la relación puede ser tan oscura que quizá sea mejor no incorporarlos desde un punto de vista de la gestión estratégica. Donde se hace presente que la adopción de la inteligencia artificial en una empresa se vea más como una inversión a largo plazo que una mera actividad operativa, aquí es donde yace la fundamentación del presente



trabajo encontrar las bases de implementación de un algoritmo de inteligencia artificial aplicado a una organización que persigue fines de lucro.

#### **1.4.4. Justificación social**

Herbert Simon, eminente científico de la administración y premio Nobel de economía, introdujo el concepto de que en la práctica es mucho más frecuente satisfacer que optimizar. Al inventar el término satisfacer como una combinación de satisfacer y optimizar, Simon describe la tendencia de los administradores a buscar una solución que sea “lo suficientemente buena” para el problema que se enfrenta. En lugar de intentar el desarrollo de una medida global de desempeño para conciliar de manera óptima los conflictos entre los diferentes objetivos deseables, incluso los criterios bien establecidos para juzgar el desempeño de los distintos segmentos de la organización, es posible utilizar un enfoque más pragmático. Las metas se pueden establecer de manera que marquen los niveles mínimos satisfactorios de desempeño en las diferentes áreas, con base quizá en niveles de desempeño anteriores o en los logros de la competencia. Si se encuentra una solución que permita que todas estas metas sean cumplidas, es posible que sea adoptada sin más requisitos. Ésta es la naturaleza de satisfacer.

Es imperativo mencionar que la construcción y la validación de un algoritmo de inteligencia artificial aplicado a la identificación de perfiles de suscripción presentado en el presente trabajo se enfoca a contribuir a la adopción de la (IA) como un medio de digitalización y adopción estratégica en un entorno digital para empresas, emprendedores y personas que incursionen en el e-commerce, negocio y/o emprendimiento con la finalidad de realizar sus actividades (actividad gerencial) con herramientas cualitativas y cuantitativas puestas en práctica en diferentes rubros o sectores. Que en consecuencia repercute en el desarrollo de las diferentes unidades productivas.



## **1.5. Alcances de la investigación**

### **1.5.1. Ámbito geográfico**

En el desarrollo de la presente investigación es imperativo definir que los elementos que serán sujetos y variables de tratamiento están en función del campo de la educación virtual que está alojado en la World Wide Web (WWW), donde el sitio web es una estructura de información y/o comunicación generada en el nuevo ámbito o espacio de comunicación (Internet), creado por la aplicación de las tecnologías de la información (tecnologías de creación, mantenimiento y desarrollo de los sitios web), que posee dos elementos fundamentales (acciones de los sujetos y contenidos) y en donde se plantean un conjunto de prestaciones que los usuarios que visitan dicho sitio web pueden ejercitar para satisfacer una o varias necesidades que posean.

### **1.5.2. Ámbito político, social o económico**

Se debe encontrar un equilibrio entre las técnicas de comunicaciones en línea y fuera de línea con base en las fortalezas y debilidades de las diferentes opciones de medios desarrollaron un marco útil para examinar las características de los medios que influyen en las decisiones sobre el comercio electrónico. Los medios fuera de línea suelen ser superiores para generar atención, estimular la atención y obtener credibilidad. Los medios en línea tienden a ser mejores para la atracción debido a la personalización, la interacción y el apoyo de boca a oreja. A menudo también se pueden ofertar en línea los productos que se pueden comprar en línea. Sin embargo, hay límites en el número de personas a las que se puede llegar a través de los medios en línea (un límite para el número de personas que buscan términos en particular) y el costo no necesariamente es siempre más bajo en los mercados competitivos.

El presente trabajo se desarrolla considerando que la inteligencia artificial en los últimos años ha estado presente dentro de las estrategias de campañas



publicitarias de Google Ads. Sin embargo, el papel que desempeña no es más que el de un medio de publicidad y no como una potente herramienta de contacto con los usuarios, conocimiento del target, autoevaluación de acciones y sus resultados como una base financiera para definir presupuestos, en este sentido la asignación de recursos y la definición concreta de resultados previsibles hacen que el beneficio de adquirir conocimientos a través de la educación virtual.

### **1.5.3. Ámbito industrial/sectorial**

(Agrawal, Etal 2001) sugieren que los sitios de comercio electrónico deben centrarse en segmentos estrechos que hayan demostrado su preferencia por un modelo de negocio. Creen que las técnicas de promoción como los clientes potenciales afiliados con sitios estrechamente dirigidos y campañas de correo electrónico dirigidas a segmentos agrupados por historiales de compra y características demográficas tienen de 10 a 15 veces más probabilidades que los banners publicitarios a los portales genéricos de atraer prospectos que hagan acciones o clics para realizar una compra. Por otra parte, los anuncios de pago por clic en Google pueden tener un mayor índice de éxito.

Es imperativo que el desarrollo del presente trabajo de investigación de optimización de recursos está directamente ligada a la comercialización de cursos virtuales, oferta de productos educativos (libros) digitales y servicios de asesoría empresarial a través del e-commerce que están presentes en los sitios de educación virtual.

### **1.5.4. Ámbito financiero**

Decidir sobre el gasto óptimo en diferentes técnicas de comunicación será un enfoque iterativo ya que se deben analizar resultados anteriores y realizar los ajustes correspondientes. Los profesionales del marketing pueden analizar la proporción del presupuesto promocional que se dedica a diferentes canales y luego compararla con la contribución de la compra de clientes que se



generaron utilizando el canal original. Este tipo de análisis, reportado por Hoffman y Novak (2000).

Uno de los objetivos del presente estudio se fundamenta en la optimización de recursos asignados a campañas digitales basadas principalmente en google ads y RRSS en academia online Rodrigo Avalos (<https://rodrigo-avalos.com/>) que en primera instancia brinda de forma especializada productos educativos (herramientas, videos, hojas de trabajo), cuya finalidad radica en el desarrollo competitivo de las empresas, negocios y emprendimientos en los tópicos de mayor impacto como ser: marketing, gestión de marca, comercialización, desarrollo del pensamiento gerencial.



## CAPITULO II

### MARCO TEÓRICO

#### 2.1. Modelaje de inteligencia artificial

La inteligencia artificial (Artificial Intelligence, o IA) es la simulación de procesos de inteligencia humana por parte de máquinas, especialmente en sistemas informáticos. Estos procesos incluyen el aprendizaje (la adquisición de información y reglas para el uso de la información), el razonamiento (usando las reglas para llegar a conclusiones aproximadas o definitivas) y la autocorrección. Hoy en día, es un término general que abarca todo, desde la automatización de procesos robóticos hasta la robótica actual.

Ha ganado prominencia recientemente debido, en parte, a los grandes volúmenes de datos, o al aumento de velocidad, tamaño y variedad de datos que las empresas están recopilando. La (IA) puede realizar tareas tales como identificar patrones en los datos de manera más eficiente que los seres humanos, lo que permite a las empresas obtener más información sobre sus datos.

El modelaje de un algoritmo que incidirá en la toma de decisiones está en función a la **experimentación** de un modelo funcional que brindará información clave y valiosa (Bronw, 1998), en la presente investigación concretamente acerca de los perfiles potenciales de nuevos suscriptores en la academia de educación virtual (<https://rodrigo-avalos.com/>).

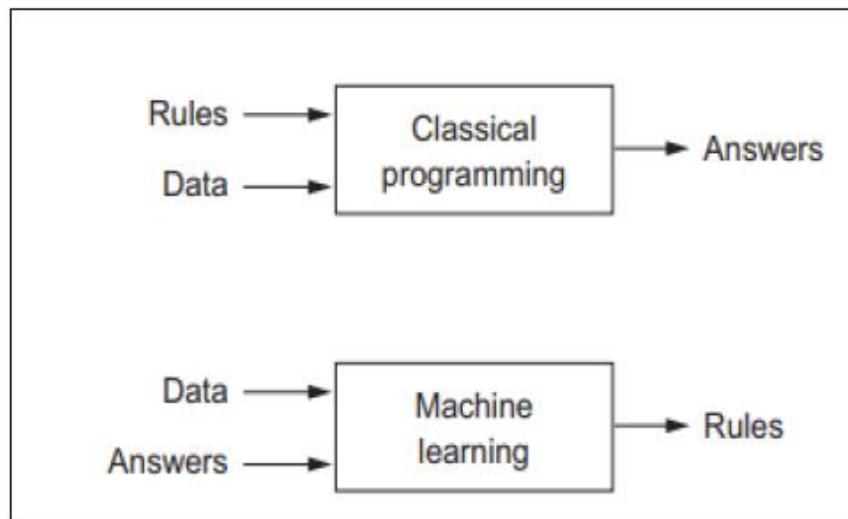
##### 2.1.1. Aprendizaje automático (Machine Learning)

Según Hurwitz (2018) afirma que, El aprendizaje automático se ha convertido en uno de los temas más importantes dentro de las asociaciones de avance que buscan enfoques imaginativos para utilizar las ventajas de la información para ayudar a la organización a obtener otro grado de comprensión. ¿Por qué añadir la inteligencia artificial a la mezcla? Con los modelos de inteligencia

artificial, las asociaciones pueden prever constantemente cambios en el negocio para que puedan anticipar lo más pronto posible.

El aprendizaje automático (AA) o machine learning se identifica como una rama de la (IA) la cual se creó inicialmente para que pudieran responder preguntas como: “¿puede un computador ir más allá de lo que le ordenamos como hacer y aprender por sí mismo como realizar una tarea específica?, ¿podría un computador automáticamente aprender reglas directamente de los datos que le pasamos?”. Con estas preguntas se inicia la creación de una nueva forma de realizar la programación, cambiando la forma clásica donde se le entrega al computador las reglas y los datos para obtener respuestas, con la AA se le entrega al computador datos y las respuestas que se obtiene para que el identifique patrones y genere sus propias reglas que serán aplicadas a los nuevos datos para obtener las respuestas esperadas. En la Figura 2, se puede observar las dos formas de programar un computador y las respuestas que se obtienen.

**Figura 2.** El nuevo paradigma de programación: *Aprendizaje automático*



**Fuente:** Bishop, Christopher (2008) Pattern Recognition and Machine Learning.





### **2.1.2. Aprendizaje supervisado**

Los algoritmos de aprendizaje supervisado están basados en un modelo predictivo, el cual está compuesto por dos grupos de datos uno para realizar el entrenamiento, otro de prueba y un mecanismo que permita evaluar si el algoritmo está haciendo las cosas bien. El conjunto de datos previamente etiquetado y clasificado, del cual ya se sabe a qué grupo, valor o categoría pertenecen son los datos de entrenamiento y son utilizados para realiza el ajuste al modelo, los datos son utilizados por el algoritmo para ir “aprendiendo” a clasificar las muestras realizando una comparación del resultado obtenido por el modelo, y el valor inicial de la muestra, realizando compensaciones con respecto al modelo y a cada error en la estimación del resultado.

Un ejemplo, de este tipo de aprendizajes son los vehículos autónomos. Y las redes neuronales artificiales (Artificial neural networks) son uno de los algoritmos que se utilizan en un enfoque de automatización. El desempeño del aprendizaje supervisado se refiere a que: la variable de información (x) se utiliza para interactuar con la variable de rendimiento (y) utilizando un cálculo. Toda la información, rendimiento, cálculo y etapa están siendo dados por personas (Van Loon, 2018). La aplicación de este tipo de aprendizaje supervisado en la identificación de perfiles de suscripciones en la educación virtual genera una ventaja en la administración de la información a la hora de establecer una estrategia de asignación de recursos publicitarios en el marketing digital.

### **2.1.3. Aprendizaje no supervisado**

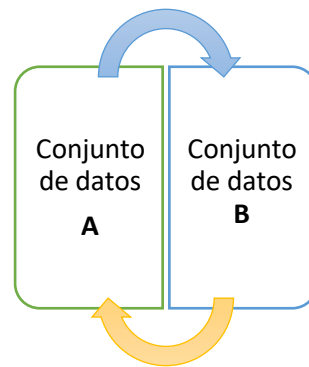
A diferencia del modelo anterior este solo tiene en cuenta para realizar sus ajustes al modelo predictivo, los datos de entrada, sin importar si están o no clasificados o etiquetados, estas características no son necesarias para realizar el entrenamiento del modelo un ejemplo de este modelo es la clasificación de imágenes y el algoritmo utilizado es K-medias (K-means). En

el aprendizaje no supervisado se puede ver que las etiquetas de salida no son entregadas a los algoritmos, sino que tienen que ser identificadas. Aunque todo el procedimiento parece ser alucinante, el aprendizaje sin ayuda puede traducirse y descubrir respuestas para una cantidad ilimitada de información, a través de la información y el componente de justificación binaria que se muestra en todos los marcos de PC. El marco no tiene referencia de información por cualquier medio (Van Loon, 2018).

#### 2.1.4. Constatación del problema

Los parámetros necesarios para iniciar a modelar un algoritmo de inteligencia artificial están dados por:

**Figura 3.** Parámetros de análisis de conjunto de datos



**Fuente:** Elaboración propia

El planteamiento del problema para ser resuelto mediante la búsqueda como se mencionó, la solución de un problema se puede llevar a cabo mediante métodos de búsqueda, para lo cual se requiere abstraer los elementos, las entidades y la forma en que van a ser manipulados por un agente que se conocerá como solucionador del problema. En segundo lugar, la función objetivo o a optimizar Variables / Parámetros que intervienen establecen las restricciones y la forma de la solución. Si sólo se desea obtener un valor final o la serie de pasos para llegar a ella.



- Elementos que se deben identificar de un problema:
- Elementos para solucionar un problema
- Elementos para solucionar un problema mediante búsqueda
- Función Objetivo Meta
- Variables / Parámetros Elementos
- Valores de las variables Valores / Estados
- Restricciones Cambiar de un estado a otro
- Forma de la solución
- La solución
- Instancia del problema Datos
- Estrategia Algoritmo de Búsqueda
- Representación del problema de acuerdo a la estrategia seleccionada
- Grafo

El proceso de identificación – representación/abstracción – obtención de datos – estrategia y solución del problema, puede ser llevado a cabo por entes que tengan un nivel de inteligencia suficiente para llevar a cabo dicho proceso, por lo cual es un tema básico en la Inteligencia Artificial. Este proceso de resolución de problemas requiere de un conjunto de datos iniciales que utilizando un conjunto de procedimientos seleccionados a priori es capaz de determinar el conjunto de pasos o elementos que nos llevan a lo que denominaremos una solución.

Cada problema que se pueda plantear es diferente, sin embargo, todos tienen ciertas características comunes que nos permiten en primera instancia clasificarlos y estructurarlos (Molina, 2009).

En segunda instancia esta estructuración, la cual se debe llevar a cabo de una manera formal, nos puede permitir resolverlos de manera automática, utilizando la representación adecuada. Dicha representación por necesidad de



uniformidad y estandarización deberá utilizar un lenguaje común para ser utilizado por los entes inteligentes.

Si abstraemos los elementos que describen un problema podemos identificar elementos como:

1. Un punto de partida, los elementos que definen las características del problema.
2. Un objetivo a alcanzar, qué queremos obtener con la resolución.
3. Acciones a nuestra disposición para resolver el problema, de qué herramientas disponemos para manipular los elementos del problema.
4. Restricciones sobre el objetivo, qué características debe tener la solución

### 2.1.5. Conjunto de datos de entrenamiento

Los datos a utilizar en el inicio del modelaje de aprendizaje supervisado están dados por la matriz [A] estructurada a continuación:

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
92	99	1	8	15	67	74	51	58	40
98	80	7	14	16	73	55	57	64	41
4	81	88	20	22	54	56	63	70	47
85	87	19	21	3	60	62	69	71	28
86	93	25	2	9	61	68	75	52	34
17	24	76	83	90	42	49	26	33	65
23	5	82	89	91	48	30	32	39	66
79	6	13	95	97	29	31	38	45	72
10	12	94	96	78	35	37	44	46	53
11	18	100	77	84	36	43	50	27	59

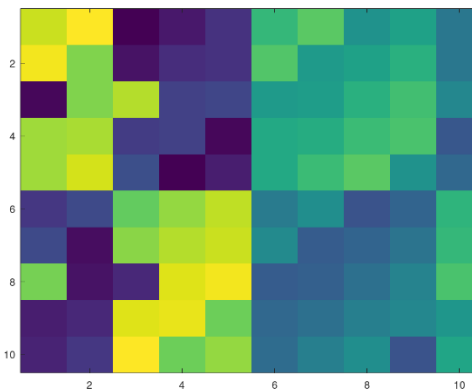
**Fuente:** Google analytics, plugin analytics parameters wordpress/Rodrigo-avalos.com

Conjunto de entrenamiento está dado por las variables:

- X1=** Cantidad de segundos página principal.
- X2=** Cantidad de segundos página portafolio.
- X3=** Cantidad de segundos página cursos virtuales.
- X4=** Cantidad de segundos página cursos argumentos prácticos.
- X5=** Cantidad de clics página principal.
- X6=** Cantidad de clics página portafolio.
- X7=** Cantidad de clics página cursos virtuales.
- X8=** Cantidad de clics página argumentos prácticos.
- X9=** Cantidad de clics wish list cursos virtuales.
- X10=** Cantidad de clics carrito de cursos.

La definición y representación de esta matriz está dada por la siguiente figura:

**Figura 4.** Visualización de (IA) matriz de confusión



**Fuente:** Elaboración propia

## 2.2. MODELACIÓN DE LA FUNCIÓN OBJETIVO

### 2.2.1. Función de costos (minimización)

El objetivo en el aprendizaje automático es extraer una relación a partir de los datos. En tareas de regresión, esta relación toma la forma de una función  $y = f(x)$ , donde  $y \in \mathbb{R}$  es una cantidad que puede ser predicho a partir de una entrada  $x \in \mathbb{R}^t$ , que por el momento debe considerarse como una colección de



medidas numéricas. La verdadera relación  $f$  nos es desconocida, y nuestro objetivo es recuperarla tanto como podamos a partir de los datos (Nasiriany, 2019).

Nuestro producto final es una función  $\hat{y} = h(x)$ , llamada hipótesis, que debe aproximarse  $f$ . Suponemos que tenemos acceso a un conjunto de datos  $D = \{(x_i, y_i)\}_{i=1}^n$ , donde cada par  $(x_i; y_i)$  es un ejemplo del mapeo de entrada -salida para ser aprendido. Dado que el aprendizaje de funciones arbitrarias es intratable, nos limitamos a algunas hipótesis clase  $H$  de funciones permitidas. Más específicamente, empleamos un parámetro paramétrico modelo, lo que significa que hay algún vector de dimensión finita  $w \in R^t$ , cuyos elementos son conocidos como parámetros o pesos, que controla el comportamiento de la función.

Es decir,  $h_w(x) = g(x, w)$  para alguna otra función  $g$ . La clase de hipótesis es entonces el conjunto de todas las funciones inducidas por las posibles elecciones de los parámetros  $w$ :

$$H = \{h_w \mid w \in R^t\}$$

Después de designar una función de costo  $L$ , que mide qué tan deficientes son las predicciones  $\hat{y}$  de la hipótesis coincidir con la salida real  $y$ , podemos proceder a buscar los parámetros que mejor se adapten a los datos minimizando esta función:

$$w^* = \min L(w)$$

### 2.2.2. Función de mínimos cuadrados

Los mínimos cuadrados es uno de los problemas de regresión más simples, pero se comprende bien y prácticamente útil. Es un problema de regresión lineal, lo que significa que tomamos  $h_w$  como de la forma  $h_w(x) = x^T w$  donde se requiere (Autor,00):

$$y_i \approx \hat{y}_i = h_w(x_i) = x_i^T w$$

Para cada  $i = 1$ ; tomados de uno en uno. Este conjunto de ecuaciones se puede escribir en forma de matriz como:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

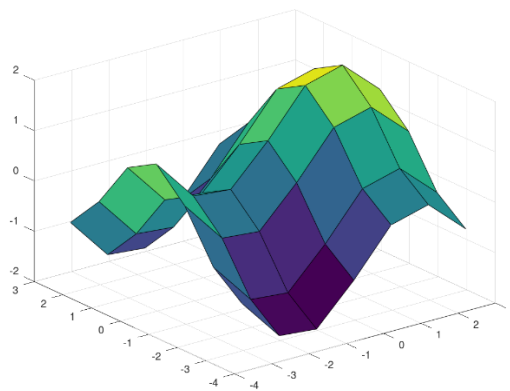
En palabras, la matriz  $x \in \mathbb{R}^{n \times d}$  tiene el punto de datos de entrada  $x_i$  como su  $i$ -ésima fila. Esta matriz es a veces llamada matriz de diseño. Normalmente  $n \geq d$ , lo que significa que hay más puntos de datos que mediciones.

En general, no habrá una solución exacta para la ecuación  $y = Xw$  (incluso si los datos fueran perfectos, considerar cuántas ecuaciones y variables hay), pero podemos encontrar una solución aproximada minimizando la suma (o equivalentemente, la media) de los errores al cuadrado:

$$L(x) = \sum_{i=1}^n (x_i^T w - y_i)^2 = \min \| Xw - y \|^2$$

Entonces, colocando los parámetros de conexión enfocada a la potencial identificación de suscripción en la educación virtual a través de la función de costos.

**Figura 5.** Visualización de la función de costos



**Fuente:** Elaboración propia.

La función de la hipótesis está dada por:

$$H_{\theta}(x) = \theta_0 + \theta_1 x$$

Donde  $h_{\Theta}(x)$  es la función hipótesis para un valor nominal  $\Theta_1$ , que a su vez es una función de  $(x)$ . Cada valor de  $\Theta_1$ , corresponde a una hipótesis diferente o una línea recta diferente de la izquierda.

### 2.2.3. Gradiente de descenso

El algoritmo gradiente de descenso para minimizar la función de costos  $J$  está dada por:

*repeat until convergence*

$$\left\{ \begin{array}{l} H_{\theta}(x) = \theta_0 + \theta_1 x \\ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad J(\theta_0, \theta_1) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \text{ for } j = 1 \text{ and } j = 0 \end{array} \right\}$$

Resulta que el gradiente de descenso es utilizado para resolver el problema de la concisión de la notación. Que está sujeta a mantener cambios pequeños para tratar de reducir  $J$  de  $\Theta_1$  hasta que converja con el mínimo local.

Desde un punto de vista ideal, los objetivos formulados deben coincidir con los de la hipótesis; sin embargo, esta coincidencia no siempre es conveniente. Muchos problemas interesan sólo a una parte de la:

$$h_{\theta}(x) = g(\theta^T x)$$

$$g(x) = \frac{1}{1 + e^{-z}}$$

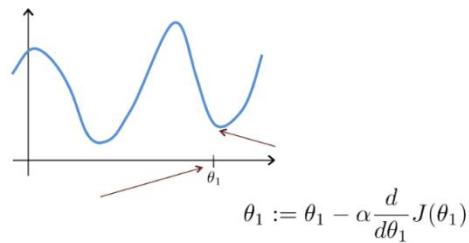
De manera que el análisis sería demasiado extenso si los objetivos fueran generales y se prestara atención especial a todos los efectos secundarios sobre el resto del conjunto de entrenamiento. En lugar de ello, los objetivos de la función  $J$  deben ser tan específicos como sea posible, siempre y cuando consideren las metas principales del tomador de decisiones y mantengan un nivel razonable de congruencia con la función objetivo.

Donde la principal causa toma forma, en un objetivo estratégico de identificar perfiles de suscripciones en la formación online, que a su vez son parámetros



de gestión propia de un prestador de servicios educativos académicos y como así también la asignación de recursos a la captación de los mismos.

**Figura 6.** Ilustración del mínimo local



**Fuente:** Elaboración propia, mínimo local.

#### 2.2.4. Regresión logística

El adjetivo a largo plazo indica que este objetivo proporciona la flexibilidad necesaria para considerar actividades que no se traducen de inmediato en ganancias, como los proyectos de investigación y desarrollo, pero que deberán hacerlo con el tiempo para que valgan la pena. La regresión logística, como la regresión lineal, se encuentra dentro de los algoritmos de aprendizaje supervisado y se trata de un modelo de regresión utilizado como método de clasificación binaria (nos ofrece una respuesta con dos valores únicamente: sí/no, hombre/mujer, 0/1, etc.), puesto que, en lugar de valores numéricos, éste permite estimar la probabilidad de que ocurra (o no) un evento como función de otro tipo de variables. Por ejemplo, puede utilizarse para determinar el sexo de una persona a partir de otras variables explicativas. En este ejemplo, el algoritmo estudia y aprende sobre un histórico de datos en el que se tiene el peso y la edad (variables explicativas) de muchas personas y si son hombre o mujer (variable de respuesta binaria). Una vez el modelo ha encontrado la relación entre las variables predictores y la variable de respuesta, puede para nuevas personas en las que se conoce el peso y la

edad predecir si son hombre o mujer. En lugar de utilizar la función de regresión lineal, este modelo se basa en una función sigmoide o logística (con forma de S) que nos permite establecer un umbral de decisión que separe las dos clases, desarrollando y otorgando información crucial para la toma de decisiones dentro del bagaje de la gestión presupuestaria en el área de marketing.

$$\text{minimize } J(\theta_0, \theta_1)$$

Dada la función anterior que describe la regresión logística que tiene como principal objetivo el de clasificar los datos en función de los parámetros de  $\Theta$ , donde los valores son 0 o 1 dentro de un conjunto de entrenamiento.

### 2.2.5. Modificación de parámetros $\Theta_0$ y $\Theta_1$

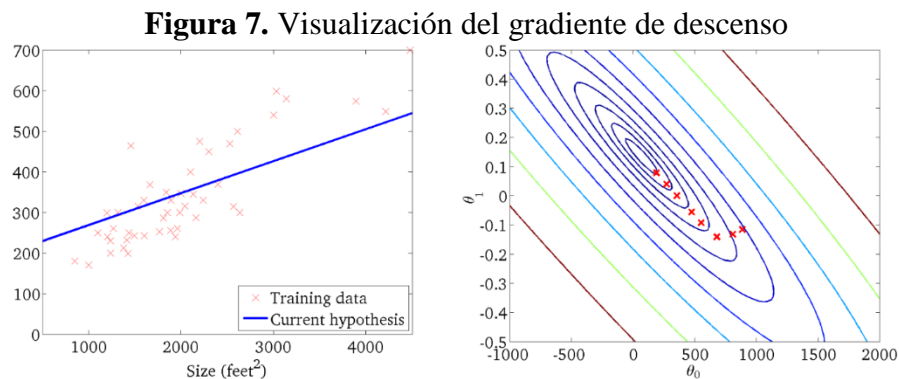
La modificación de los parámetros  $\Theta_0$  y  $\Theta_1$  serán consideradas como los hiper parámetros de validación y evaluación del modelo dado por la siguiente fórmula que establece los aspectos más importantes del aprendizaje automático, el cual establece un modelo ya sea de carácter complejo o simple en función de los datos de entrenamiento.

$$J(\theta) = -\frac{1}{2} \left[ \sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_{\theta}(x^{(i)})) \right]$$

La anterior formula está sujeta a la modificación de los parámetros de  $\Theta_0$ ,  $\Theta_1$  donde la principal variación dependerá de las magnitudes o el volumen de los datos de entrenamiento. Dados por los vectores  $\Theta$ , que serán parametrizados en función de los valores de predicciones correctas (precisión del modelo), frente a la distribución del conjunto de entrenamiento del algoritmo (precisión esperada).

### 2.2.6. Gradiente de descenso

El presente trabajo se desarrolla considerando que el algoritmo de gradiente de descenso esta generado oscilaciones en un conjunto de entrenamiento para convergencia en el óptimo global y no así en el óptimo local de un conjunto de datos dado. Sin embargo, el papel que desempeña no es más que el de un medio identificación o colocación de etiquetas de clasificación y no como una potente herramienta de contacto con los usuarios, conocimiento del target, autoevaluación de acciones y sus resultados como una base financiera para definir presupuestos enmarcados en el área de marketing.



Fuente: Elaboración propia

## 2.3. PERFILES DE CONSUMIDOR

### 2.3.1. Información de perfiles potenciales

En el desarrollo de la presente investigación es imperativo definir que los elementos que serán sujetos y variables de tratamiento están en función del campo de educación virtual que está alojado en la World Wide Web (WWW), donde el sitio web es una estructura de información y/o comunicación generada en el nuevo ámbito o espacio de comunicación (Internet), creado por la aplicación de las tecnologías de la información (tecnologías de creación, mantenimiento y desarrollo de los sitios web), que posee dos elementos fundamentales (acciones de los sujetos y contenidos) y en donde se plantean un conjunto de prestaciones que los usuarios que visitan el sitio web



(<https://rodrigo-avalos.com/>) pueden ejercitar para satisfacer una o varias necesidades que posean dentro de la educación virtual; según Kotler (2003) los clientes serán clasificados en función de su aporte a la empresa en términos de participación de mercado y las acciones que hacen dentro del punto de venta:

- **Clientes Actuales:** Son aquellos (personas, empresas u organizaciones) que le hacen compras a la empresa de forma periódica o que lo hicieron en una fecha reciente.
- **Clientes Potenciales:** Son aquellos (personas, empresas u organizaciones) que no le realizan compras a la empresa en la actualidad pero que son visualizados como posibles clientes en el futuro porque tienen la disposición necesaria, el poder de compra y la autoridad para comprar.

Ahora bien, las aplicaciones de los términos anteriormente mencionados enfocados a un sitio web de educación virtual, serán vinculados y estratificados de la siguiente manera para la modelación de la información de salida del algoritmo de inteligencia artificial adoptado en el presente trabajo de investigación:

- **Cliente A – suscriptor premium:** Este tipo de clientes es el que genera el volumen de ventas actual, por tanto, es la fuente de los ingresos que percibe el sitio web en la actualidad y posee una alta prioridad en la asignación de recursos publicitarios.
- **Cliente B – suscriptor consecuente;** está en constante aprendizaje compuesto por hábitos de tiempo y recuperación (por medio de certificaciones) asiduos en el sitio web de educación virtual.
- **Cliente C – suscriptor anzuelo;** Este tipo de clientes compra su producto o servicio de forma regular, pero basados en la cantidad de descuento que una empresa le ofrece.

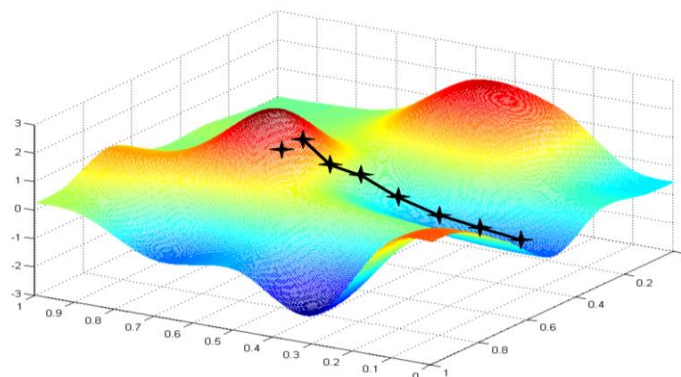
- **Cliente D – suscriptor retención;** Son aquellos (personas, empresas u organizaciones) que no le realizan compras a la empresa en la actualidad pero que son visualizados como posibles clientes en el futuro porque tienen la disposición necesaria, el poder de compra y la autoridad para comprar. Este tipo de clientes es el que podría dar lugar a un determinado volumen de ventas en el futuro (a corto, mediano o largo plazo) y, por tanto, se los puede considerar como la fuente de ingresos futuros.

En consecuencia, la determinación del tipo de perfil de suscripción en la educación virtual corresponde en esencia a la detección de comportamientos y acciones asiduos que tienen los visitantes que realizan en el sitio web (<https://rodrigo-avalos.com/>) dando así, características latentes de comportamientos que desembocan en los perfiles A, B, C y D también cabe resaltar que las suscripciones que posee el sitio de educación virtual son de pago único, tomando como base este tipo de suscripción para la modelación del algoritmo de machine learning para la determinación de una estrategia de asignación de recursos publicitarios en el marketing digital.

### 2.3.2. Actualización y asignación de perfiles

La generación de la información de los perfiles está representada a continuación:

**Figura 8.** Agrupación de perfiles de suscripción



Fuente: Elaboración propia



La gráfica nos muestra 7 diferentes grupos de datos que se muestran con agrupación considerable para la determinación de perfiles, es decir la tipología de perfiles estaría basada en estos parámetros dado que el conjunto de datos esta agrupados en las cruces presentes en el gráfico.

### **2.3.3. Recomendaciones al momento de asignar etiquetas al perfil del cliente actual**

- Decidir sobre el gasto óptimo en diferentes técnicas de comunicación será un enfoque iterativo ya que se deben analizar resultados anteriores y realizar los ajustes correspondientes.
- Las decisiones pueden analizar la proporción del presupuesto promocional que se dedica a diferentes canales y luego compararla con la contribución de la compra de clientes que se generaron utilizando el canal original.
- El objeto de estudio se fundamenta en un sitio web que en primera instancia brinda de forma orgánica productos educativos (herramientas, videos, hojas de trabajo), pero la principal fuente de ingresos es la comercialización de cursos virtuales cuya finalidad radica en el desarrollo productivo de las empresas, negocios y emprendimientos en los tópicos de mayor impacto como ser: marketing, gestión de marca, comercialización, desarrollo del pensamiento gerencial, que por su naturaleza están fuera de los alcances de la presente investigación.

### **2.3.4. Toma de decisiones individual**

Las fases que experimenta el consumidor en el proceso de compra se resume en:

**Figura 9.** Representación del proceso de compra del consumidor



**Fuente:** Elaboración propia a partir del libro Comportamiento del consumidor de Solomon (2008).

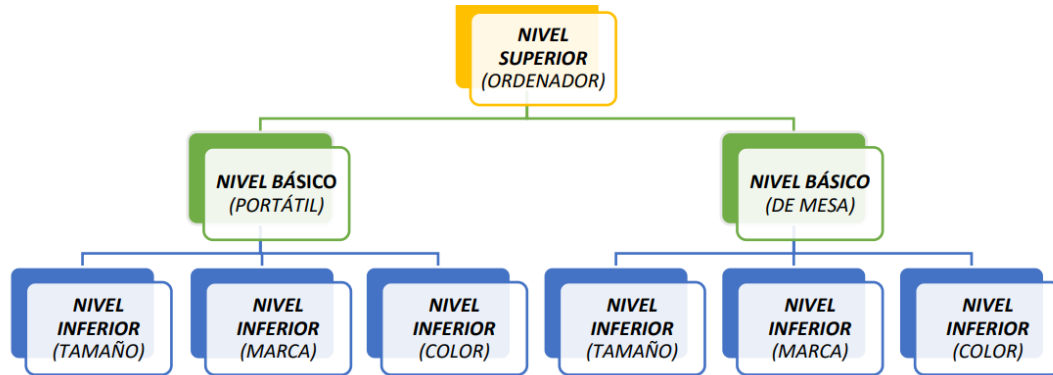
Este proceso comienza cuando el consumidor experimenta un problema, en este caso, una necesidad que siente que aún no está cubierta. Siente que hay una gran diferencia entre el estado en el que se encuentra y el que desea lograr. Este problema puede surgir de forma natural o, lo que es más común, las empresas lo crean gracias a los estudios llevados a cabo previamente por los responsables de marketing. A partir de este momento el consumidor pasa por una serie de fases hasta que lleva a cabo la compra.

Una vez haya percibido el problema, su siguiente paso es buscar información para poder solucionarlo y tomar la mejor decisión. En este caso hablaríamos de información previa a la compra, que a su vez puede dividirse entre interna (la buscamos en nuestra memoria) y externa (la completamos con información recibida de nuestro entorno). En esta fase el tiempo es mayor cuando la compra es importante y por lo tanto se pone en juego una cantidad mayor de dinero o la satisfacción de una necesidad que se considera muy importante o esencial. Entre tantos otros aspectos también afecta el riesgo que se asume con respecto al que se está dispuesto a asumir.

Una vez recogida la información necesaria se llega al punto más complicado y que más tiempo requiere, ya que cada vez son más las alternativas ante las que nos encontramos en el mercado, se evalúan las distintas alternativas con el fin de tomar la mejor decisión. En esta fase se analizan las posibilidades que el consumidor tiene de verdad en cuenta, "conjunto de consideración"; muchas alternativas se desechan directamente debido a malas experiencias previas con la marca o el precio.

Estos conjuntos de alternativas a tener en cuenta se clasifican en función de distintos criterios para facilitar este proceso.

**Figura 10.** Clasificación de productos según los niveles de abstracción



**Fuente:** Elaboración propia a partir de Solomon (2008).

Cuando se han ordenado y analizado las distintas alternativas, el consumidor debe decantarse por una de ellas. El consumidor olvida muchas de las características diferentes que ofrece el producto y se centra en un número reducido de ellas que considera importantes. La decisión se puede tomar en base a la regla de decisión no compensatoria (elimina las opciones que no ofrecen unos atributos básicos) y a la regla de decisión compensatoria (permiten que una alternativa supere sus deficiencias).

No podemos asumir que este proceso finaliza con la decisión de compra del producto ya que una vez realizada la compra el consumidor evalúa los resultados y asume si ha realizado una buena elección o, por el contrario, se equivocó en su decisión.

### 2.3.5. Influencia externa

Para comprender el modo en el que un individuo ha tomado una decisión, debemos conocer los factores externos que le influyen. Se pueden reconocer cinco factores que explicaré más detalladamente a continuación:





- *Ambiente demográfico*: Podemos comprender el modo en el que se comporta un consumidor cuando conocemos el entorno demográfico en que ha nacido y crecido.

Influyen en este el nivel de formación, número de habitantes y pirámide demográfica o nivel económico del país, por eso no podemos olvidar los cambios que se dan en este para ofrecer el producto que mejor se adapte en cada momento.

- *Ambiente económico*: Cada individuo toma decisiones teniendo en cuenta la renta nacional (valor monetario del conjunto de bienes y servicios producidos en un país) y el nivel adquisitivo (la renta de la que disponen y cómo se divide entre consumo y ahorro).

Además de la situación económica bajo la que se rige la sociedad en la que se opera, las decisiones se ven influenciadas por las características económicas de cada individuo.

- *Cultura*: La cultura es un determinante con gran peso en el comportamiento humano, influye en la toma de decisiones, comportamientos y condiciona la forma en la que se vive en cualquier sociedad.

La cultura es el conjunto de normas, creencias, valores y costumbres que son aprendidas por la sociedad, y al contrario de las características biológicas que son innatas, estas se aprenden a lo largo del ciclo de vida.

Para comprender de qué manera influye la cultura en este proceso debemos considerar que los valores son significativos, ya que los consumidores se comportarán de un modo en el que alcancen valores considerados socialmente adecuados.

- *Familia*: Es el grupo social principal que influye en la personalidad, actitudes o motivaciones del individuo. Este grupo enseña a la persona desde la infancia las pautas de consumo que debe llevar a cabo.



Pero además de tener gran influencia en la personalidad de sus miembros, según la estructura familiar o ciclo de vida de la familia, las decisiones serán distintas. El ciclo de vida familiar es una variable compuesta por la combinación de variables demográficas tales como: estado civil, tamaño, edad de los miembros o nivel de ingresos.

Tradicionalmente este ciclo comenzaba con el matrimonio, seguía con los hijos y terminaba cuando uno de los cónyuges moría. En la actualidad esta teoría no es válida, ya que se han creado nuevas estructuras familiares como familias monoparentales, monoparentales o reconstituidas.

- *Grupos sociales:* Son grupos de referencia con los que el individuo se identifica y que influyen en la formación de creencias, actitudes y comportamientos. Se puede tratar de grupos de los que uno es miembro (primarios o secundarios) o grupos a los que se pretende pertenecer (grupos de aspiración o disociación).

Estos grupos ejercen distintos tipos de influencias: informativas (los miembros intercambian información), comparativas (ofrecen los criterios de evaluación de su propia imagen) y normativas (establecen las normas de comportamiento del grupo).

Los consumidores en muchas ocasiones llevan a cabo una decisión de compra porque lo han hecho la mayoría de miembros que pertenecen a su grupo o porque creen que si compran un producto podrán integrarse en un grupo al que pretenden unirse.

Estos grupos tienen mayor influencia en la compra de algunos tipos de productos como: coches, ropa o productos tecnológicos. Algunos miembros tienen mayor influencia que otros debido a un mayor conocimiento o un estatus superior. Estos individuos son considerados como líderes de opinión y su influencia está reforzada por la credibilidad que se les atribuye.



## CAPITULO III

### METODOLOGÍA DE LA INVESTIGACIÓN

#### 3.1. Tipo de análisis

La presente investigación está enmarcada dentro del manejo de datos en su gran mayoría que a su vez es un parámetro para la identificación de los componentes necesarios de un software dirigido a la identificación de los datos históricos, que darán a lugar a la identificación de los perfiles potenciales de nuevos suscriptores, donde el tipo de análisis **correlacional** será el catalizador del esquema a utilizar para el desarrollo del contenido científico.

#### 3.2. Marco metodológico

La utilización de un algoritmo de inteligencia artificial está fundamentada desde la perspectiva **cuantitativa**, que por su naturaleza se estudiara un conjunto de datos homogéneos agrupados.

#### 3.3. Diseño de la investigación

El modelaje de un algoritmo que incidirá en la toma de decisiones está en función de un modelo funcional que brindará información acerca de los perfiles potenciales de nuevos suscriptores en una academia online, en este sentido, la presente investigación tendrá un **diseño transversal**.

Es conveniente mencionar también que se emplea el lenguaje de programación Python el cual modelará en función de sus librerías específicas, dado las condiciones del algoritmo a emplearse.

#### 3.4. Unidades de análisis

Con el propósito de apoyar la convergencia de los datos escrutados en función a los obtenidos en el empleo de instrumentos de recolección de big data y su posterior análisis dentro de un campo estrictamente compilatorio para la depuración de los datos obtenidos debe establecerse las siguientes variables

que serán parte de este estudio en sus conjeturas preliminares y no así limitantes.

**Tabla 1.** Construcción de indicadores de entrada

Variable cuantitativa → # de visitas a los sitios de educación virtual.
Variable cuantitativa → # duración de la visita.
Este parámetro de relación será evaluado como el indicador #1.
Variable cuantitativa → # de clics en los sitios de educación virtual.
Variable cuantitativa → # botones en los cuales se hicieron clic.
Este parámetro de relación será evaluado como el indicador #2.
Variable cuantitativa → # de visitas a páginas específicas dentro del sitio.
Variable cuantitativa → # de segundos de scrolling.
Este parámetro de relación será evaluado como el indicador #3.
Variable cuantitativa → # formularios de registro generados en las visitas.
Variable cuantitativa → # segundos en el formulario de registro.
Este parámetro de relación será evaluado como el indicador #4.
Variable cuantitativa → # reproducciones de los cursos.
Variable cuantitativa → # de clics en la página de compras (carro de compras).
Este parámetro de relación será evaluado como el indicador #5.

Fuente: Elaboración propia.

### 3.5. Determinación del universo y tamaño de la muestra

Siendo el objeto de análisis los datos obtenidos en una inmersión y discriminación de parámetros modelables serán puestos en función en los sitios web, utilizando la librería Numpy de python que será el componente básico de granulador de los datos procedentes del uso y tiempo que proporcionen dichos eventos granulares.

Siendo así, la determinación del universo será en consecuencia el resultado del empleo granular de un algoritmo en específico dado por:



```
>>> from sklearn import linear_model
>>> X = [[0., 0.], [1., 1.], [2., 2.], [3., 3.]]
>>> Y = [0., 1., 2., 3.]
>>> reg = linear_model.BayesianRidge()
>>> reg.fit(X, Y)
BayesianRidge()
```

En función de los parámetros:

```
>>> reg.predict([[1, 0.]])
array([0.50000013])

>>> reg.coef_
array([0.49999993, 0.49999993])
```

Una vez obtenido el conjunto de datos apto para iniciar el proceso de escogencia de la técnica de Machine Learning que se usará para desarrollar el algoritmo de árboles de decisión objeto de esta investigación, se realizó un tercer análisis teniendo en cuenta el número de variables y el número de ejemplos recolectados. Nuestro conjunto de datos está formado por 36 *variables* y 22205 *observaciones*, como se observa nuestro conjunto de datos tiene una alta dimensionalidad y esto implica comprometer la eficiencia del clasificador escogido por tener un modelo con una complejidad alta que nos podría llevar a un overfitting, además se debe tener en cuenta que, por el número de variables obtenidas, se puede estar corriendo el riesgo de tener atributos ruidosos que pueden tener el mismo peso que los atributos relevantes. Como solución a este problema se utilizó la técnica Análisis de Componentes Principales que nos permitió reducir la dimensionalidad del conjunto de datos preservando en lo posible, la aleatoriedad de los datos del conjunto original.



### 3.6. Hipótesis de estudio

La hipótesis está comprendida por:

*Un algoritmo de machine learning basado en arboles de decisión genera la identificación de perfiles de suscripción en la educación virtual permitiendo definir una estrategia de asignación de recursos publicitarios enfocados a las campañas de marketing digital.*

La definición de las variables en su sentido académico tendrá un enfoque desarrollado por la academia de educación online (<https://rodrigo-avalos.com/>) que formaran parte de la presente investigación, que, de hecho, son variables que se verán depuradas con carácter de empleo en las librerías de inteligencia artificial que posee Python. En efecto la hipótesis tendrá una **incidencia alta** en relación a las variables que la componen según el modelo que se adoptará que en este caso es el algoritmo de clasificación denominado árboles de decisión.

### 3.7. Indicadores de estudio

**Precisión**, mide cuantos casos que han sido clasificados en una clase pertenecen realmente a esa clase, dada por:

$$\text{Precisión} = \frac{vp}{vp + fp}$$

Donde:

- $vp$ : Verdadero positivo, individuos de la clase que son clasificados como tales.
- $fp$ : Falso positivo, individuos de la clase contraria a la que son clasificados con respecto a esta clase.
- $vn$ : Verdadero negativo, individuos que el predictor a predicho que pertenecen a la clase objetivo, pero que no pertenecen a ella.



- $f_n$ : Falso negativo, individuos que el predictor a predicho que no pertenecen a la clase objetivo.

Todas estas métricas son válidas cuando la variable de salida escogida es discreta. Cuando la variable es continua se han de aplicar otras métricas diferentes, como “least squares”, porque no se espera que el modelo acierte exactamente la salida. Ahora la precisión hace que el modelo de machine learning empleado en el presente trabajo sea evaluado para ver la consistencia de la asignación de recursos publicitarios en el marketing digital con un enfoque en la reducción de costos e incremento en el retorno de la inversión publicitaria.

**Entropía**, El termino entropía fue usado por primera vez por Clausius en 1864 e introducido en la teoría de la información por Shannon en 1948. La entropía de Shannon se define para un sistema con N estados posibles donde  $p_i$  es la probabilidad de encontrar el sistema en el estado  $i$  –ésimo. En sí, la entropía nos ayuda a analizar la información que se utiliza para medir la cantidad de información útil. Se calcula como:

$$H(S) = - \sum_{x \in X} p(x) * \log_2 p(x)$$

Nos indicará la cantidad de observaciones empleadas para el modelado e implementación del algoritmo de árboles de decisión, así entonces dando paso a la sustracción o ampliación de variables de entrada que en nuestro caso específico son las acciones del visitante en el sitio web de educación virtual.

**Gini Impurity**, es la medida de cuan a menudo se puede clasificar mal un elemento si fuera clasificado de forma aleatoria de acuerdo con la distribución de las clases del subconjunto, se calcula de la siguiente manera:

$$i(t) = \sum_{i,j} C(i|j) * p(i|t) * p(j|t)$$



Donde:

- $i(t)$  es la impureza de la clase  $t$ .
- $C(i, j)$  son los subconjuntos de clases de  $i, j$ .
- $p(i|t)$  la proporción de elementos de la clase  $t$  sobre  $i$ .
- $p(j|t)$  la proporción de elementos de la clase  $t$  sobre  $j$ .

El criterio de división de Gini es el decremento de la impureza definida como:

$$\Delta i(s, t) = i(t) - pL i(tL) - pR i(tR)$$

Donde  $pL$  y  $pR$  son las probabilidades de escoger el hijo derecho o el hijo izquierdo, cada uno se estima como:

$$pL = p(tL)/p(t), pR = p(tR)/p(t)$$

Esta impureza nos permite trabajar la calidad de cada nodo para generar una evaluación del modelo más específico en cada rama del árbol de decisiones generado, haciendo que las etiquetas o variables sean tomadas en mayor o menor medida para el nivel de predicción y optimización del trabajo del algoritmo. Dando paso a la mejor calidad de cada rama del árbol de decisión que tiene como principal función la identificación de perfiles de suscripción en la educación virtual.



### 3.8. Operacionalización de variables

**Tabla 2.** Operacionalización de las variables de estudio

Variable	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Instrumento
Algoritmo de Machine learning basado en arboles de decisión.	Se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos. Utilizando el discriminante que agrupa en cada una de sus ramas los resultados más próximos y no así los mejores.	Agrupar los datos representativos en ramificaciones concretas de consecución que reflejan la disposición de obtención de valor.	1. Machine learning	<ul style="list-style-type: none"> <li>• Nivel de precisión en la predicción (tasa de aciertos).</li> <li>• Entropía</li> <li>• Gini impurity</li> </ul>	Software libre Python
Perfiles de suscripción	Son las características definidas en el uso de una categoría de servicio en un momento previsual.	Clasifica la tendencia de su comportamiento con fines de asignación eficiente de los recursos de marketing.	2. Identificación de perfiles de consumidores.	<ol style="list-style-type: none"> <li>1. Plan de promoción cliente tipo A (Premium).</li> <li>2. Plan de promoción cliente tipo B (Consecuente).</li> <li>3. Plan de promoción cliente tipo C (Anzuelo).</li> <li>4. Plan de promoción cliente tipo D (Retención).</li> </ol>	Dataset compuesto por las características y acciones del visitante en el sitio web de educación virtual ( <a href="https://rodrigo-avalos.com/">https://rodrigo-avalos.com/</a> ), (El DataFrame preparado y elaborado por el tesista).

**Fuente:** Elaboración propia.



### 3.9. Unidades de análisis

Definidos ya los parámetros de la investigación es posible definir una estructura de trabajo en función de los requerimientos de la educación virtual, en este caso otorgada por los sitios Udemy, BSG Institute y el sitio rodrigo-avalos. Se empleará el siguiente procedimiento de selección, transformación y depuración de data:

- Normalización de la data: Se estandarizan las variables a utilizar.
- Agregación de la data: Se crean las tablas donde se almacenará la información para su posterior análisis.
- Generación del DataFrame: Se crean los campos calculados y las etiquetas a emplearse.
- Filtrado de la data: En este punto se filtra solo los campos que brindan valor al modelo.
- Selección de atributos: Se definen los atributos que serán analizados en la minería de datos.
- Data Mining: En esta Fase se utiliza el algoritmo elegido (Árboles de Decisiones) aplicado al modelo desarrollado previamente con la data recopilada y transformada.

## CAPITULO IV

### MARCO PRÁCTICO

#### 4.1. Análisis de campo

Se ha recopilado datos sobre un conjunto de visitantes en un sitio web empleando el script:

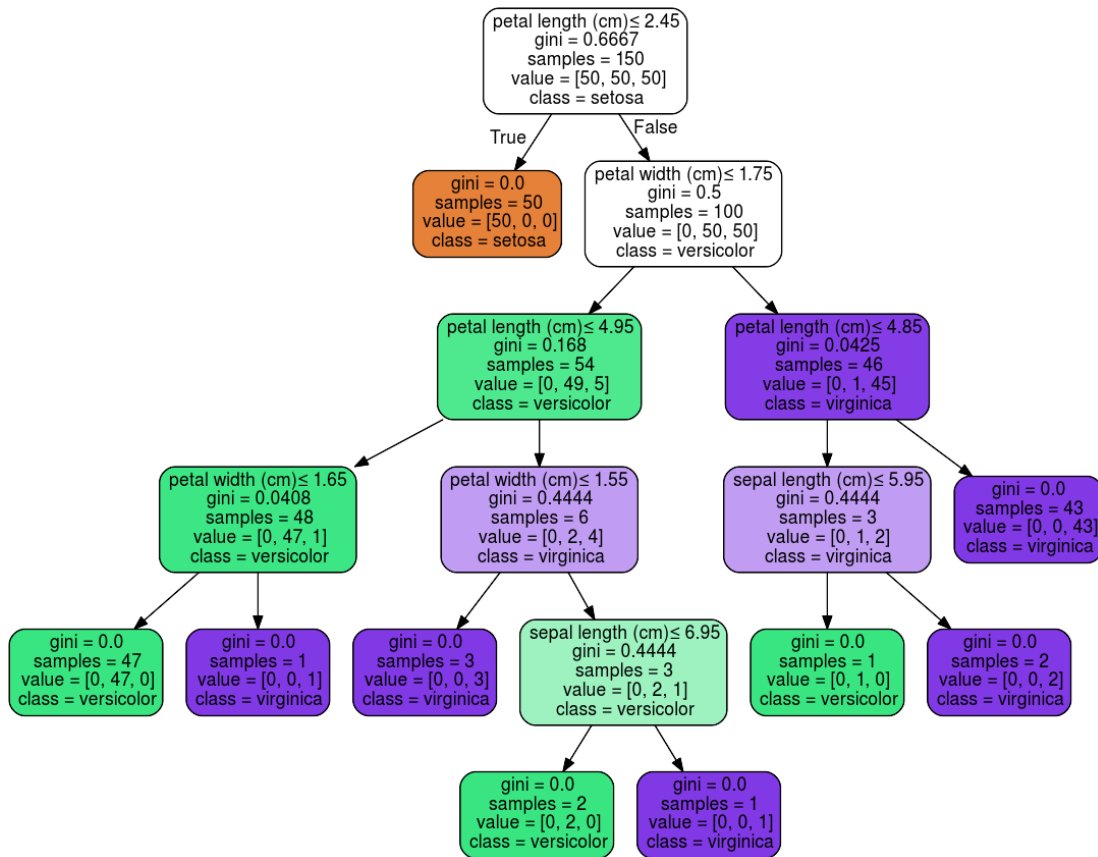
```
$ pip3 install requests
import requests

URL = 
page = requests.get(URL)
```

Durante la recopilación de tratamiento los datos se transformaron en una muestra de clasificadores binarios, y puede usar la parte de entrenamiento del conjunto de datos para construir un árbol de decisiones y luego usarlo para predecir la clase de un paciente desconocido. En esencia, tomar una decisión sobre qué tipo de visitante es y cual promoción tiene mayor probabilidad de incidir en el proceso de compra.

Los árboles de decisión se construyen dividiendo el conjunto de entrenamiento en distintos nodos, donde un nodo contiene toda o la mayor parte de una categoría de datos. Si miramos el diagrama aquí, podemos ver que es un clasificador de características.

**Figura 11.** Árbol de decisión en base a la clasificación de etiquetas



**Fuente:** Elaboración propia en base a Liberia IRIS, Python (2006).

Entonces, como se mencionó, queremos identificar al visitante de un sitio web y entregarle un programa de comunicación que estará en función de sus características (etiquetas) de desempeño en un sitio web. Partiendo de esta premisa se puede construir un árbol de decisiones considerando los atributos uno por uno. Primero, se elige un atributo de nuestro conjunto de datos. Calculando la importancia del atributo en la división de los datos. A continuación, la división de los datos según el valor del mejor atributo, se trabajará cada rama repetidamente para el resto de los atributos. Después de construir el árbol de decisiones, se podrá emplear para predecir la clase de casos desconocidos; o en nuestro caso, la promoción adecuada para un nuevo visitante en función de sus características.

### 4.1.1. Análisis de las variables

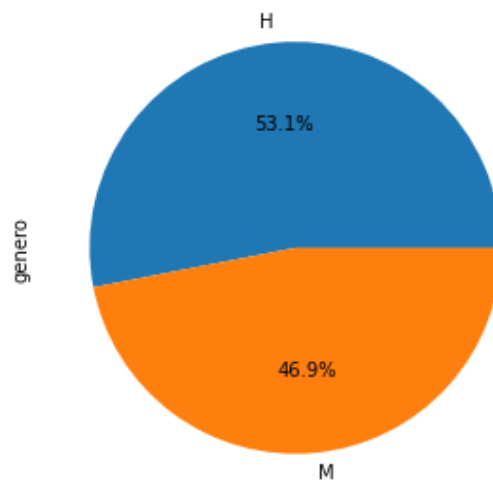
Iniciando la estructura del análisis inicial compuesto por los descriptivos pertenecientes se obtuvo las características de la data (DataFrame) correspondiente al análisis descriptivo de los perfiles con una muestra igual a 571 observaciones.

**Figura 12.** Datos categóricos - género

```
[28] perfil.genero.value_counts()
↳ H    303
   M    268
   Name: genero, dtype: int64
```

**Fuente:** Elaboración propia.

**Figura 13.** Porcentajes correspondientes - género



**Fuente:** Elaboración propia.

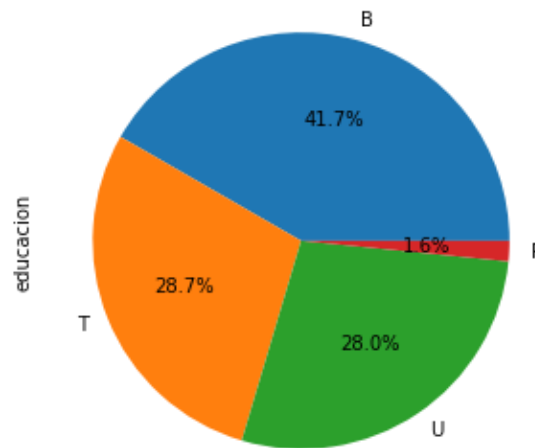
En la figura 13 se puede apreciar que el género predominante en las visitas a sitios web de educación virtual corresponde a los hombres con un 53.1% equivalente a 303 observaciones frente al 46.9% de mujeres que está compuesto por 268 féminas.

**Figura 14.** Datos categóricos - nivel educativo

```
▶ perfil.educacion.value_counts()
↳ B    238
   T    164
   U    160
   P     9
   Name: educacion, dtype: int64
```

**Fuente:** Elaboración propia.

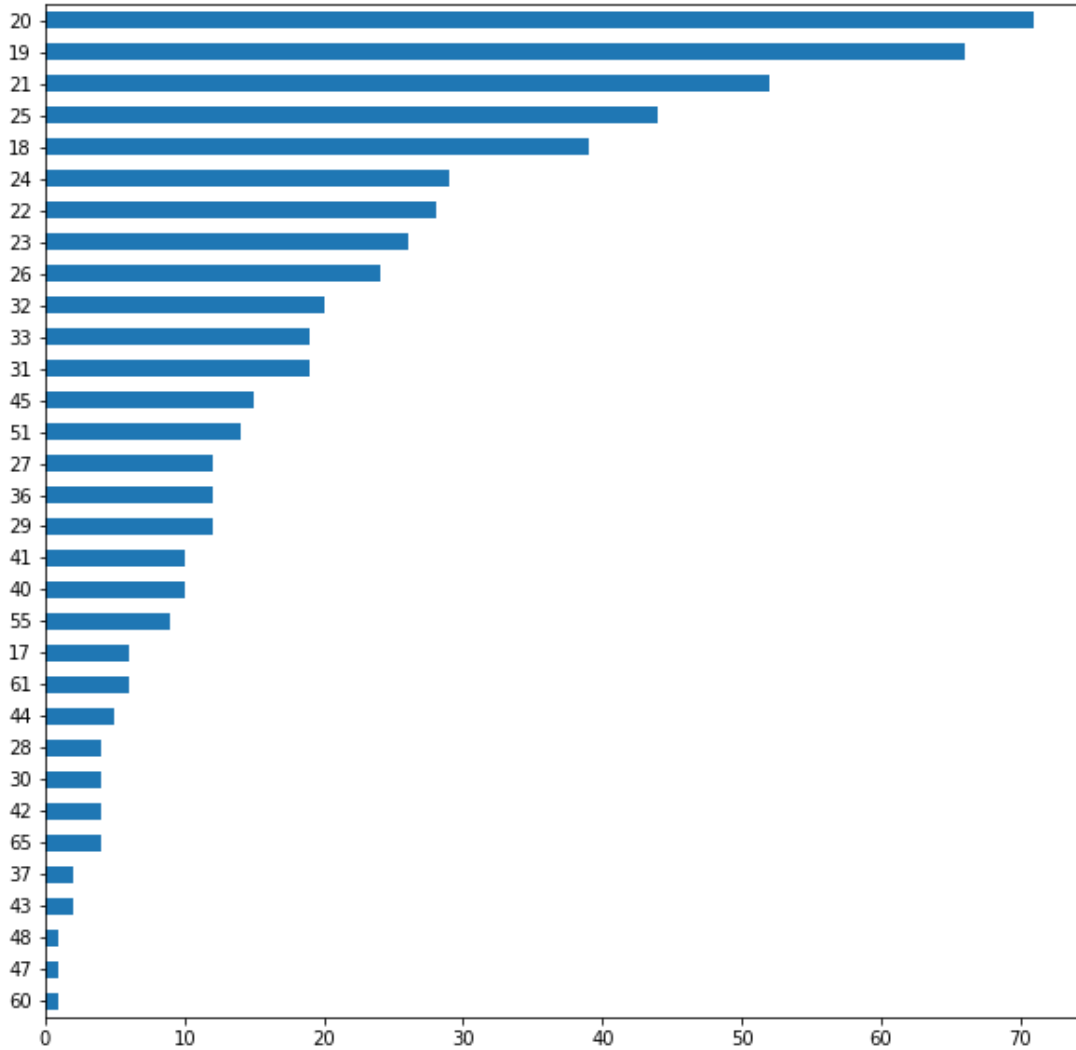
**Figura 15.** Porcentajes correspondientes – nivel educativo



**Fuente:** Elaboración propia.

Correspondientes a las categorías B = bachiller con 41.7%, nivel técnico con 28.7% y nivel universitario con 28% donde P = 1.6% corresponde al 1.6% que no define su ocupación se puede inferir que la preponderancia de las personas que buscan cursos online pertenece a jóvenes que acaban de salir del nivel secundario.

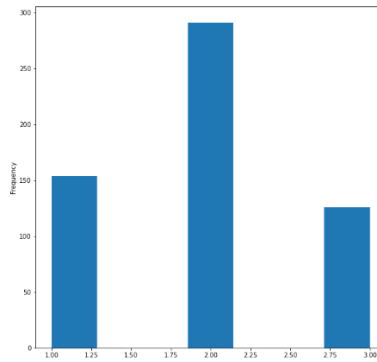
**Figura 16.** Frecuencia - etiqueta edad



**Fuente:** Elaboración propia.

Observando la concentración de las edades se con 71 observaciones que tiene la característica principal (en función de la edad) o de observación enmarcada en la edad de los visitantes en un sitio web de educación correspondiente a los 20 años, validando con certeza la primera inferencia en la categoría de la educación.

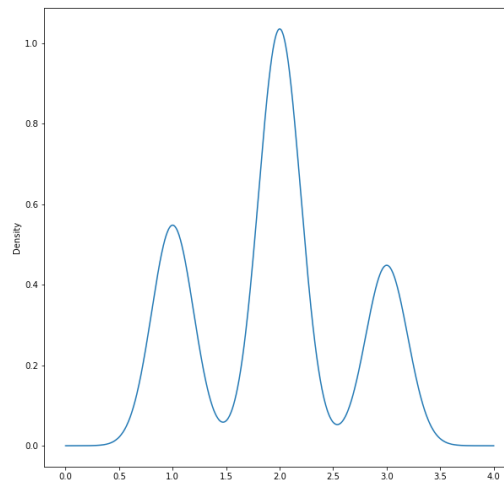
**Figura 17.** Distribución - etiqueta ciudad



**Fuente:** Elaboración propia.

La concentración de las conexiones categorizadas en 1= ciudades capitales, 2 = ciudades intermedias, 3 = ciudades pequeñas están definidas en la categoría 2 y 1 con 291 y 154 observaciones las cuales concentran un 77%, empleando una distribución empleando la librería matplotlib de Python podemos encontrar también la probabilidad de densidad reflejada en la figura 9 con sus correspondientes probabilidades de categorización.

**Figure 18.** Probabilidad de distribución



**Fuente:** Elaboración propia.



#### 4.1.2. Constatación del problema en base a las variables validadas

En esta fase, se seleccionó las variables consideradas para el modelo y luego se ingresó a la herramienta (Data Tools de Microsoft) para que la data sea procesada por el algoritmo seleccionado (Árboles de Decisión) y posteriormente validada (Cross Validation 10-Fold).

Para ello se hace un análisis de las variables que afectan al modelo y se decidió seleccionar las siguientes con los siguientes tipos de datos presentados en la herramienta:

**Tabla 3.** Variables de modelado

<b>Variable</b>	<b>Tipo</b>
Tipo de compra	<i>Input</i>
Monto de compra	<i>Input</i>
Número de productos	<i>Input</i>
Estado civil	<i>Input</i>
Género	<i>Input</i>
Edad	<i>Input</i>
Fecha de compra	<i>Input</i>
Lugar de compra	<i>Input</i>
Tipo de cliente A	<i>Predict Only</i>
Tipo de cliente B	<i>Predict Only</i>
Tipo de cliente C	<i>Predict Only</i>
Tipo de cliente D	<i>Predict Only</i>

**Fuente:** Elaboración propia.

Luego de ejecutar la herramienta se genera la validación de los resultados que sirve para saber el porcentaje error que tiene el modelo Para ello se utiliza la matriz de confusión.

**Tabla 4.** Matriz de confusión

Variable	Descripción	Predicción	Predicho	Real
TP	Verdadero positivo	Correcta	Desaprobado	Desaprobado
TN	Verdadero positivo	Correcta	Aprobado	Aprobado
FP	Falso positivo	Incorrecta	Desaprobado	Aprobado
FN	Falso negativo	Incorrecta	Aprobado	Desaprobado

Fuente: Elaboración propia.

Además, se utilizó la matriz de confusión para interpretar los resultados obtenidos luego de ejecutar la técnica de validación (Cross Validation 10-Fold) y así, obtener el porcentaje de error del modelo en la primera ejecución. A continuación, se presenta los resultados:

Formula Porcentaje de Error

$$\%Error = 1 - \frac{TP + TN}{TP + TN + FP + FN}$$

El siguiente cuadro muestra el nivel de ejecución:

**Tabla 5.** Nivel de conteo y predicción

Diagnóstico	Conteo	Nivel de predicción
TP	1645	17.64%
TN	4561	22.55%
FP	8655	45.81%
FN	1446	8.96%

Fuente: Elaboración propia.

## 4.2. Resultados

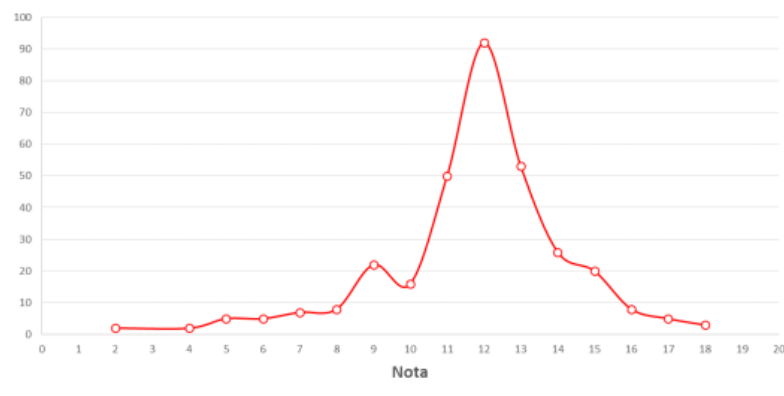
El resultado obtenido es:

**Tabla 6.** Resultados de entrenamiento

%Error	6.88%
% Aciertos	89.62%

Fuente: Elaboración propia.

**Figura 19.** Nivel de errores

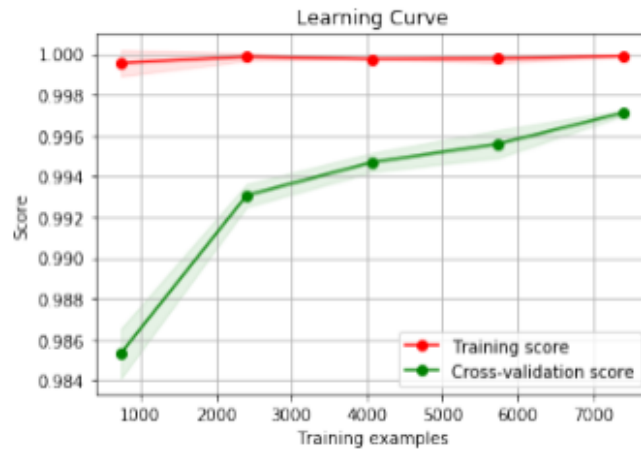


Fuente: Elaboración propia

En el anterior gráfico se observa que la mayoría de los errores encontrados en el modelo de entrenamiento versus el resultado real se concentra entre los ciclos 10 y 14. Esto se debe a que entre dicho intervalo se encuentra el género. También, se puede observar que el punto máximo se encuentra en el estado civil, 0.5 puntos menos que la variable género.

Dicho gráfico es un sustento de que el modelo está siendo eficiente y consecuente con la información brindada por la estandarización de los datos.

**Figura 20.** Nivel de aprendizaje del algoritmo



**Fuente:** Elaboración propia.

Esta prueba la realizamos para poder determinar con cuál criterio (entropía o gini) el algoritmo Random Forest detecta si la consulta realizada es o no SQL injection. Para lo cual se realizó pruebas de forma independiente para el criterio gini y entropía, para cada uno se realizó un total de 7 pruebas.

El tamaño de las muestras es igual  $n = 7$ , los datos que se recogen es el porcentaje de predicción de forma general con criterio Gini y Entropía.

La hipótesis estadística es:

- $H_0$  = La clasificación de consultas es alta con el criterio Gini.
- $H_1$  = La clasificación de consulta se realizar de igual manera tanto con Gini y Entropía.
- El nivel confianza de 95%.

**Tabla 7.** Nivel de Gini, entropía y diferencia

<i>Muestra</i>	<i>Gini</i>	<i>Entropía</i>	<i>Diferencia</i>	<i>TP/TN</i>
1	99.4654645	99.412144	0,05332050	-4
2	99.54654644	99.42135	0,12519644	-2.5
3	99.896652	99.546444	0,35020800	1.2
4	99.44644	99.321544	0,12489600	1.5
5	99.1154	99.01264	0,10276000	-5.1
6	99.2144	99.1321544	0,08224560	-4.3
7	99.365	99.21544	0,14956000	2.5

**Fuente:** Elaboración propia.

Donde la suma de los negativos es en valor absoluto es  $T^- = 22$  y para valores positivos es  $T^+ = 6$ , el valor más pequeño  $T$  es valor estadístico de prueba.

El valor crítico para la estadística de prueba  $T$  para  $\alpha = 0.05$  y  $n = 6$  porque en la muestra 3 la diferencia da 0. De la tabla de Wilcoxon Signed-Ranks obtenemos  $T_{crit} = 0$ ;  $T_{crit} = 0 < 3$ , como  $T$  es mayor al valor crítico afirmamos que la  $H_0$  es verdadera. Es decir, la clasificación de si una consulta es mala o legítima se realiza mejor bajo el criterio Gini.

En este proceso puede ser tedioso, pues puede exigir conocimiento de otras áreas relacionadas con la matemática, estadística y el área de seguridad informática. Considerando que para el tratamiento de datos en aprendizaje automático debe ser numérico, entonces hacemos el proceso de la transformación. Empezamos por dividir la consulta, con una herramienta llamada SqlParser desarrollado en Python. La herramienta nos permitirá identificar los tipos de cadena existentes en una consulta. Esta herramienta nos permitirá identificar si una cadena es single, Float, DDL, Keyboard, paréntesis, identificador de puntuación. Estos nuevos parámetros generados nos permitirán en la siguiente etapa mejorar nuestro tratamiento de datos.

Inicialmente se utilizó el análisis de componentes principales para reducir la dimensionalidad del conjunto de datos destinado para entrenamiento. Para el proceso de entrenamiento y validación se aplicaron diferentes técnicas de machine



Learning asignando diferentes valores de parámetros en cada una de ellas, luego se evaluó su desempeño utilizando las medidas precisión, recall y F1. El mejor F1 con un 85% se obtuvo con las Redes Neuronales con 250 neuronas en la capa oculta, factor de regularización=0.1 y 1000 iteraciones aplicados al conjunto de datos con 74 componentes principales.

También se tuvo en cuenta al análisis ROC, mostrado en la figura 20, en este se observa un desplazamiento de la curva hacia arriba y hacia a la izquierda, tendiendo a una probabilidad de clasificación correcta del 85% con las redes neuronales, el cual fue el clasificador que muestra mayor área bajo la curva.

Por último, tenemos el siguiente código del algoritmo de árboles de decisión:

```
import pandas as pd
perfiles = pd.read_csv('/content/drive/My Drive/2020 - AI -
ROHAB/nomarlizados.perfiles2020.csv')
perfiles.sample (10)

print("Dimensiones del dataset: \n")
print(perfiles.shape, '\n')
print("Tipo de datos: \n")
print(perfiles.dtypes)

from sklearn.model_selection import train_test_split
X = perfiles.iloc[:, [0,1,2,3,4,5,6]].values
y = perfiles.iloc[:, 7].values
X_train , X_test, y_train, y_test = train_test_split(X, y , test_size
= 0.25, random_state = 2)

from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train )

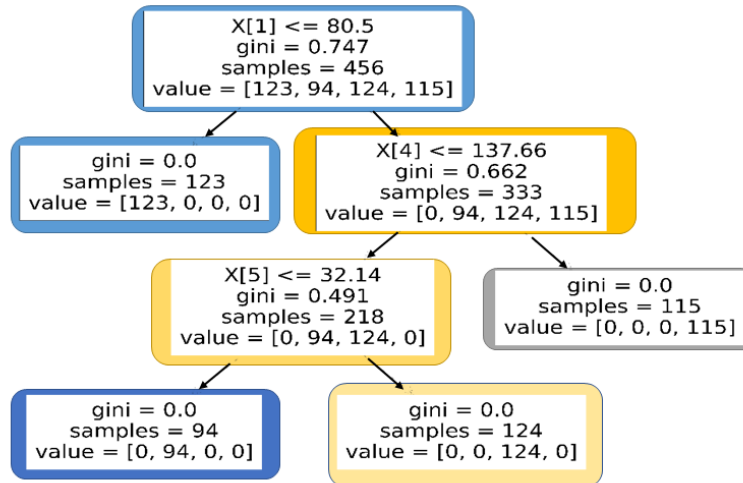
y_pred = clf.predict(X_test)
y_pred

score = clf.score(X_test, y_test)
print('Precisión: %0.4f' % (score))

from sklearn.tree import export_graphviz
import graphviz
from sklearn import tree
from sklearn.tree import plot_tree
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 200,150
rcParams['figure.figsize'] = 200,150
plot_tree(clf)
```

En consecuencia, al digitar el código en Python se logra obtener la siguiente representación:

**Figura 21.** Visualización del árbol de decisión aplicado



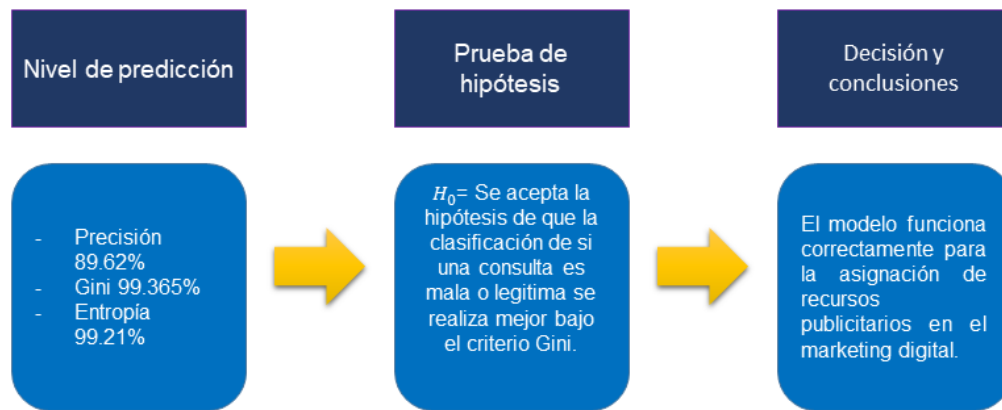
**Fuente:** Elaboración propia.

Definidos los parámetros es posible concluir que el árbol de decisiones para la clasificación de perfiles suscriptores tiene suficientes categorías para mostrarnos el nivel de carteleras o sea clientes en nuestra base, empero, la diferencia de la entropía hace que los parámetros deben ser analizados por el valor menor 0.0574 o inferior para mejorar el desempeño de los niveles de cada rama.

### 4.3. Comprobación de hipótesis

La generación de un modelo válido empleando un algoritmo de inteligencia artificial en base a un árbol de decisiones para identificar los tipos de clientes y conociendo los resultados anteriores lo podemos resumir en el siguiente gráfico:

**Figura 22.** Resumen de la validación del algoritmo de inteligencia artificial de árboles de decisión



**Fuente:** Elaboración propia

Otorgando en primera instancia un instrumento funcional al momento de emplear los recursos destinados en el marketing digital con la capacidad de identificar el tipo de perfil que visitan el sitio web (<https://rodrigo-avalos.com/>) con un grado de precisión del 89.62%.

Empero la disposición de todo el entramado de los datos, donde su aplicabilidad emerge de la optimización de los recursos en campañas digitales para obtener un mayor retorno de la inversión y reducción de costes, está en función de que nuestra hipótesis definida por:

*Un algoritmo de machine learning basado en arboles de decisión genera la identificación de perfiles de suscripción en la educación virtual permitiendo definir una estrategia de asignación de recursos publicitarios enfocados a las campañas de marketing digital.*

Permitiendo que se pueda discernir el impacto que posee un algoritmo de inteligencia artificial para la creación de estrategias de asignación de recursos publicitarios, por tanto, la información preliminar antes de emplear el algoritmo se muestra en el siguiente gráfico:



**Figura 23.** Situación de la asignación de recursos publicitarios y el ROI antes de la adopción de un algoritmo de inteligencia artificial para la identificación de perfiles en la educación virtual



Fuente: Elaboración propia

Es pieza clave al momento de la transición basada en información descriptiva y unidimensional a emplear modelos más complejos que coadyuvan a la toma de decisiones en el ámbito del marketing operativo y asignación de recursos publicitarios y en concreto arguyendo a una entidad que se especializa en la comercialización online de cursos virtuales, este caso el sitio web (<https://rodrigo-avalos.com/>), es menester entender que *la hipótesis y sus resultados tienen una incidencia es alta y a su vez implica un relacionamiento positivo tanto a nivel operativo como administrativo*. En conclusión, la hipótesis se valida dado que el algoritmo de machine learning basado en una estructura de árboles de decisión, aporta información valiosa acerca del tipo de perfiles de suscripción en la educación virtual dando paso a la creación de una estrategia de asignación de recursos publicitarios en campañas de marketing digital en la cual su incidencia se estima en un incremento promedio del **ROI en 17.14%** y una reducción de **costos promedio del 5.85%** lo cual repercute en una asignación más eficiente y un retorno mayor de la inversión publicitaria según el siguiente cuadro:

**Figura 24.** Impactos de la implantación de un algoritmo basado en árboles de decisión en la creación de estrategias de asignación de recursos en el marketing digital



**Fuente:** Elaboración propia.

Acotando a la información financiera puede resumirse a continuación teniendo como base un estimado de que el gasto promedio por mes en una campaña digital es de Bs. 6385 empleado exclusivamente en RRSS como: Facebook, Instagram y Google ADS.

**Tabla 9.** Gasto promedio publicitario gestión 2019

Sistema publicitario digital	Inversión publicitaria promedio (expresado en bolivianos)
Facebook, Instagram Ads	4386.25
Google Ads	1998.75
<b>Total</b>	<b>6385.00</b>

**Fuente:** Elaboración propia.

Cabe resaltar que la inversión publicitaria ha tenido como principal característica la construcción de diversas campañas para diferentes públicos objetivos, dado que la aplicación del algoritmo de Machine learning nos permite identificar la proporción correspondiente a los diferentes tipos de perfiles – cliente que ingresan al sitio web significando una reducción de costes promedio mensual de Bs. 373.5



## CAPITULO V

### PROPUESTA

#### 5.1. PROPUESTA – MODELADO Y EJECUCIÓN DE UN MODELO DE MACHINE LEARNING APLICADO A LA IDENTIFICACIÓN DE PERFILES DE SUSCRIPCIÓN EN LA EDUCACIÓN VIRTUAL EN UNA ORGANIZACIÓN CON FINES DE LUCRO

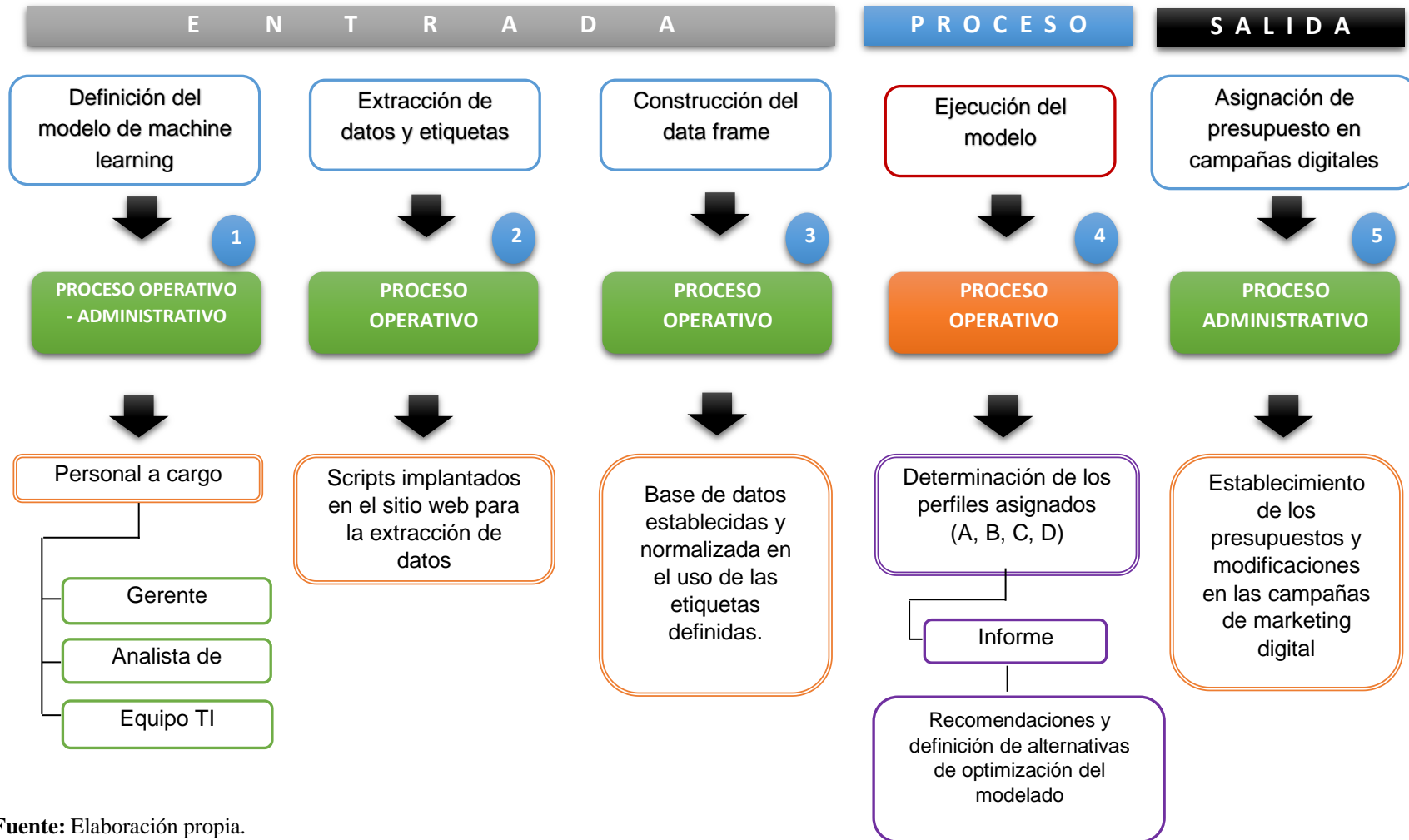
##### 5.1.1. Justificación

La propuesta es fundamentada principalmente por cuatro pilares; el primero, proviene de la observación del número y etiquetas de cada visitante en el sitio web basadas en la normalización de dichos datos en un DataFrame manipulable en Python estos datos provienen del sitio de educación virtual [www.rodrigo-avalos.com](http://www.rodrigo-avalos.com), el segundo pilar se basa en la identificación y aplicación de un algoritmo de Machine learning para determinar la posibilidad de generar compras de cursos online lo cual se enmarca en asignar los recursos en las campañas digitales concernientes a los perfiles identificados descritos en el primer pilar. Tercer pilar, estrechamente relacionado con el segundo, es el número cursos vendidos aplicando el modelo de inteligencia artificial. El último pilar, señala los resultados de una secuencia correcta de pasos al implementar el algoritmo de Machine learning basados en árboles de decisión con objeto de retroalimentar tanto las dimensiones del modelo como de sus parámetros de actualización para la adopción en una organización con fines de lucro.

##### 5.1.2. Objetivo general

- Desarrollar un modelo de Machine learning basado en árboles de decisión enfocado en la identificación de perfiles de suscripción en la educación online aplicando un enfoque sistémico como soporte al marketing digital.

**Figura 25.** Propuesta de adopción del aprendizaje automático en una organización con fines de lucro



Fuente: Elaboración propia.

Bajo el precepto de que el proceso operativo – administrativo 1 y el proceso operativo 2 se dieron prioridad en la construcción del presente trabajo ahora se enfocará en los procesos operativos 3, 4 y el proceso administrativo 5 entonces así definimos en concordancia con los anteriormente mencionado; ahora se presenta un proceso que tiene como campo de acción al gerente como líder en la implementación de la digitalización y desarrollo de habilidades en inteligencia artificial como principal herramienta de toma de decisiones en una organización que se dedica en la comercialización de cursos virtuales.

### 5.1.3. Propuesta de programa de identificación de suscriptores en una academia de educación virtual

El siguiente código correspondiente al lenguaje de programación Python permitirá generar un árbol de decisión funcional y enfocado a la identificación de perfiles de suscripción en un sitio web de educación virtual, siendo el código fuente el que se presenta a continuación:

**Figura 26.** Propuesta de programa de identificación de suscriptores en la educación virtual (código fuente para el lenguaje de programación Python)

<pre>import pandas as pd perfiles = pd.read_csv('/content/drive/My Drive/2020 - AI - ROHAB/nomaralizados.perfiles2020.csv') perfiles.sample (10)  print("Dimensiones del dataset: \n") print(perfiles.shape, '\n') print("Tipo de datos: \n") print(perfiles.dtypes)  from sklearn.model_selection import train_test_split X = perfiles.iloc[:, [0,1,2,3,4,5,6]].values y = perfiles.iloc[:, 7].values X_train , X_test, y_train, y_test = train_test_split(X, y , test_size = 0.25, random_state = 2)  from sklearn import tree clf = tree.DecisionTreeClassifier() clf = clf.fit(X_train, y_train )  y_pred = clf.predict(X_test) y_pred score = clf.score(X_test, y_test) print('Precisión: %0.4f' % (score))  from sklearn.tree import export_graphviz import graphviz from sklearn import tree from sklearn.tree import plot_tree from matplotlib.pylab import rcParams rcParams['figure.figsize'] = 200,150 rcParams['figure.figsize'] = 200,150 plot_tree(clf)</pre>	<p>1. Importando el data frame.</p> <p>2. Mostrando etiquetas del data frame.</p> <p>3. Definiendo las variables de entrada y salida del modelo.</p> <p>4. Importando el modelo de árboles de decisión.</p> <p>5. Ejecución del modelo (predicción).</p> <p>6. Visualización del árbol de decisión.</p>
---	---

Fuente: Elaboración propia



En concreto la empleabilidad del programa descrito con anterioridad se la describirá a continuación paso por paso la cual es parte del proceso operativo 3

## 5.2. PROCESO OPERATIVO 3

En la siguiente propuesta, se establece una secuencia de pasos para implementar un algoritmo de aprendizaje automático (Machine L) aplicado en la determinación de variables clave en la identificación de perfiles de suscripción en el campo de la educación virtual, el árbol de decisiones que se utilizará será un algoritmo de clasificación para construir un modelo a partir de datos históricos de visitantes a un sitio web de enseñanza y sus acciones dentro de un sitio web. Luego, se emplea el árbol de decisiones capacitado para predecir la clase de visitante y para encontrar una promoción adecuada para un nuevo visitante.

### 5.2.1. Importando el DataFrame

Importando las siguientes librerías:

**Tabla 8.** Importar librerías a emplear en Python

Acción 1:	Responsable	Código en Python
Importando las siguientes librerías:	Analista de datos	<pre>import pandas as pd perfiles = pd.read_csv('/content/drive/My Drive/2020 - AI - ROHAB/nomarlizados.perfiles2020.csv') perfiles.sample (10)</pre>

**Fuente:** Elaboración Propia

### 5.2.2. Visualización del conjunto de datos (dataset)

Los datos que se recopila para este estudio. Se los ha recopilado datos sobre un conjunto visitantes, todos los cuales ingresan al sitio web de enseñanza (<https://rodrigo-avalos.com/>). Durante el curso de su visita, cada visitante tuvo las siguientes variables de sus acciones dentro de la plataforma, tipo de visitante, gasto promedio, numero de clics en el botón comprar, horario tiempo promedio (en segundos), scrolling en el sitio, días y tipo de promoción que visualizó y generó una conversión.

Según el siguiente resumen su naturaleza y características se puede apreciar:

**Tabla 9.** Visualización de los datos correspondientes al archivo entrenamiento 2020

```
import pandas as pd
perfiles = pd.read_csv('/content/drive/My Drive/2020 - AI - ROHAB/entrenamiento12020.csv', sep=',')
perfiles.head()
```

	tipo_visitante	gasto_prom	numero_clicks_comprar	horario	tiempo_promedio	scrowling	días	promocion
0	1	137	89	3	13.32	50.80	6	4
1	1	42	89	1	49.39	36.48	1	3
2	2	153	21	2	19.06	17.91	4	2
3	1	75	62	1	129.95	40.88	2	4
4	4	188	36	1	9.27	59.17	4	3

**Fuente:** Elaboración Propia.

Características adicionales:

**Tabla 10.** Descriptivos del dataframe

```
print("Dimensiones del dataset: \n")
print(perfiles.shape, '\n')
print("Tipo de datos: \n")
print(perfiles.dtypes)
```

Dimensiones del dataset:  
(571, 8)

Tipo de datos:

tipo_visitante	int64
gasto_prom	int64
numero_clicks_comprar	int64
horario	int64
tiempo_promedio	float64
scrowling	float64
días	int64
promocion	int64
dtype:	object

**Fuente:** Elaboración Propia

La esencia de este trabajo consiste en crear un modelo para averiguar qué promoción podría ser apropiado para un visitante en el sitio web. Los conjuntos de características de este conjunto de datos anteriormente descritos nos brindarán información adecuada para que cada promoción se asigne de manera directa con las acciones realizadas en su visita.

Es una muestra de clasificador binario, y puede usar la parte de entrenamiento del conjunto de datos para construir un árbol de decisiones y luego usarlo para predecir



la clase de visitante que ingresa o el tipo de promoción que se muestra por cada visualización.

### 5.3. PROCESO OPERATIVO 4

#### 5.3.1. Declarar las variables de trabajo

Usando entrenamiento.csv como los datos de perfiles2020.csv leídos por panda, se procede a declarar las siguientes variables:

X = como la matriz de características (datos de entrenamiento.csv)

Y = como vector de respuesta (objetivo)

A continuación, se elimina la columna que contiene el nombre del objetivo, ya que no contiene valores numéricos.

#### 5.3.2. Configurar el árbol de decisiones

Usaremos la división de entrenamiento / prueba en nuestro árbol de decisiones. En el libro de Python a importar:

**Tabla 11.** Importando el algoritmo de árbol de decisión

Acción 2:	Responsable	Código en Python
Importando el árbol de decisiones.	Analista de datos.	<pre>[ ]: from sklearn.model_selection import train_test_split     ...</pre>

Fuente: Elaboración Propia

Ahora train\_test\_split devolverá 4 parámetros diferentes. Los nombraremos:

X\_trainset, X\_testset, y\_trainset, y\_testset

El train\_test\_split necesitará los parámetros:

X, y, test\_size = 0.3 y random\_state = 3.

X e y son las matrices requeridas antes de la división, test\_size representa la proporción del conjunto de datos de prueba y random\_state asegura que obtengamos las mismas divisiones.



**Tabla 12.** Visualización de los elementos y variables en el modelo de árboles de decisión

Acción 3:	Responsable	Código en Python
Definiendo el conjunto de entrenamiento y el test de prueba.	Analista de datos	<pre> from sklearn.model_selection import train_test_split X = perfiles.iloc[:, [0,1,2,3,4,5,6]].values y = perfiles.iloc[:, 7].values X_train , X_test, y_train, y_test = train_test_split(X, y , test_size = 0.2, random_state = 2) </pre>

Fuente: Elaboración Propia

### 5.3.3. Modelado

Primero crearemos una instancia de DecisionTreeClassifier llamada drugTree. Dentro del clasificador, especifique criterio = "entropía" para que podamos ver la ganancia de información de cada nodo.

**Tabla 13.** Modelado del árbol de decisiones

Acción 4:	Código en Python
Modelado y ejecución del modelo de clasificación.	<pre> import numpy as np from sklearn.model_selection import GridSearchCV from sklearn.tree import DecisionTreeClassifier  grid_param = {'max_depth':np.arange(2,20), 'max_leaf_nodes':np.arange(2,20)}  dt = DecisionTreeClassifier(random_state=42)  grid = GridSearchCV(dt, grid_param , n_jobs=-1, cv=5, iid=False)  grid.fit(X_train,y_train) print(grid.best_params_) print("") print(grid.best_estimator_) </pre>
Responsable:	Analista de datos

Fuente: Elaboración Propia

### 5.3.4. Predicción

Se realiza predicciones en el conjunto de datos de prueba y posteriormente serán almacenadas en una variable llamada predTree.

**Tabla 14.** Determinación de los niveles y parámetros en la ejecución del modelo

<b>Acción 5:</b>	Código en Python
Predicción	<pre>DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',                         max_depth=9, max_features=None, max_leaf_nodes=17,                         min_impurity_decrease=0.0, min_impurity_split=None,                         min_samples_leaf=1, min_samples_split=2,                         min_weight_fraction_leaf=0.0, presort='deprecated',                         random_state=42, splitter='best')</pre>
<b>Responsable:</b>	Analista de datos

Fuente: Elaboración Propia

### 5.3.5. Evaluación

Nuestro modelo tiene un desempeño de aprendizaje del 99.13% donde está por encima del 50% que estima un desempeño excelente en la clasificación del tipo de visitante al sitio web (<https://rodrigo-avalos.com/>).

**Tabla 15.** Evaluación del modelo

<b>Acción 6:</b>	Código en Python
Evaluación del modelo	<pre>[15] y_pred = clf.predict(X_test)  y_pred array([1, 3, 1, 4, 2, 3, 3, 2, 3, 1, 3, 4, 3, 3, 4, 3, 3, 4, 1, 3, 4, 3, 2,        3, 3, 1, 1, 1, 1, 4, 2, 3, 4, 2, 3, 4, 2, 3, 4, 2, 4, 1, 4, 1, 2,        1, 1, 1, 2, 2, 1, 2, 1, 2, 3, 3, 3, 4, 4, 4, 1, 3, 1, 4, 4, 2, 4,        4, 4, 1, 4, 1, 4, 3, 4, 3, 4, 4, 4, 3, 1, 1, 2, 4, 3, 1, 3, 3, 4,        2, 3, 1, 1, 2, 2, 3, 4, 1, 2, 1, 2, 3, 4, 1, 2, 1, 1, 4, 1, 2, 1,        1, 4, 4, 3, 1])  [18] score = clf.score(X_test, y_test)  print('Precisión: %0.4f' % (score)) Precisión: 0.9913</pre>
<b>Responsable:</b>	Analista de datos

Fuente: Elaboración propia.

### 5.3.6. Visualización

A continuación, se debe colocar el siguiente código, previo a la visualización del árbol decisión que se entrenó:

```
from sklearn.tree import export_graphviz
import graphviz
from sklearn import tree
from sklearn.tree import plot_tree
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 200,150
```

Esto nos permitirá tener la siguiente visualización expuesta en la tabla 17:

**Tabla 16.** Visualización del árbol de decisión construido empleando el lenguaje de programación Python.

<b>Acción 7:</b>	Código en Python
Visualización:	<pre> graph TD     Node0["X[1] &lt;= 80.5 gini = 0.747 samples = 456 value = [123, 94, 124, 115]"]     Node1["gini = 0.0 samples = 123 value = [123, 0, 0, 0]"]     Node2["X[4] &lt;= 137.66 gini = 0.662 samples = 333 value = [0, 94, 124, 115]"]     Node3["X[5] &lt;= 32.14 gini = 0.491 samples = 218 value = [0, 94, 124, 0]"]     Node4["gini = 0.0 samples = 115 value = [0, 0, 0, 115]"]     Node5["gini = 0.0 samples = 94 value = [0, 94, 0, 0]"]     Node6["gini = 0.0 samples = 124 value = [0, 0, 124, 0]"]      Node0 --&gt; Node1     Node0 --&gt; Node2     Node2 --&gt; Node3     Node2 --&gt; Node4     Node3 --&gt; Node5     Node3 --&gt; Node6     </pre>
<b>Responsable:</b>	Analista de datos

Fuente: Elaboración Propia

### 5.3.7. Interpretación

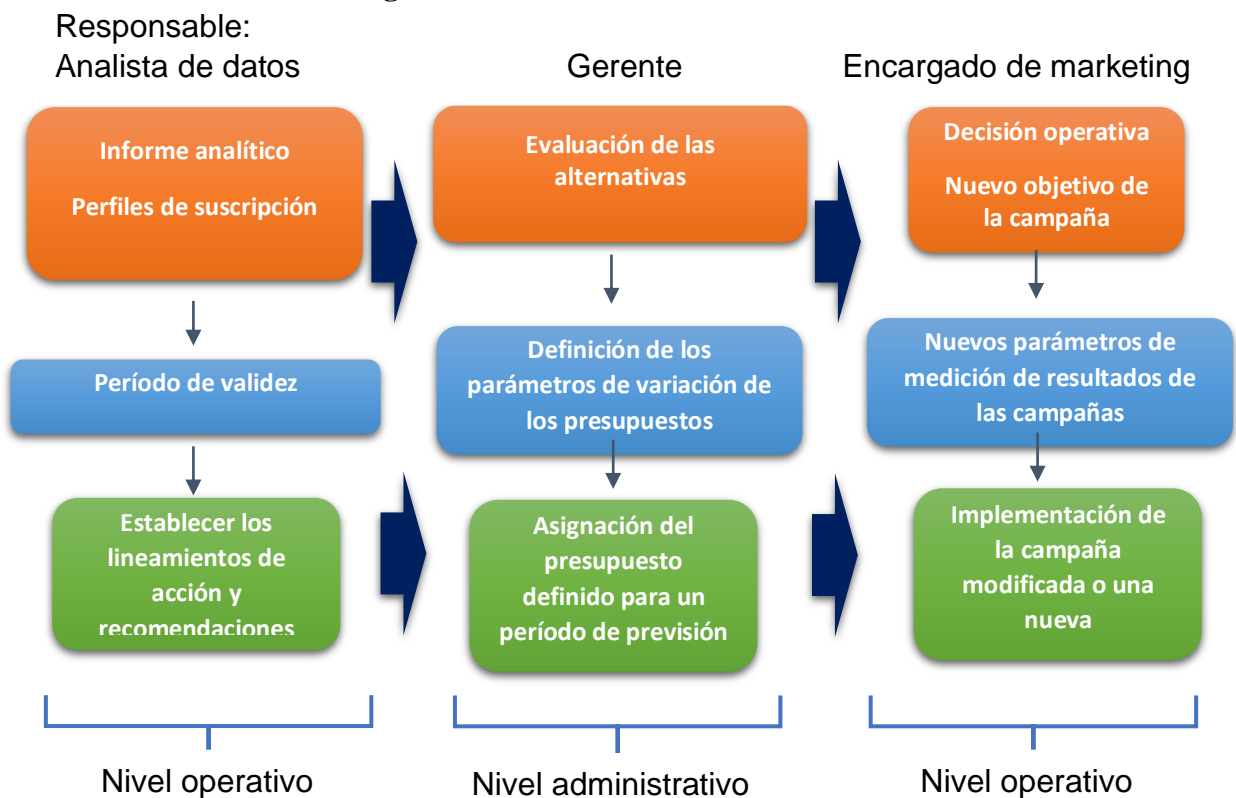
Al ser X1 la variable gasto promedio (expresado en dólares), existen 2 eventos, el primero (gini = 0.0) con un número de 123 casos de los cuales son los que no tienen un gasto real menor a 80.5\$ dólares son estos los que deben ser descartados en las visualizaciones de promociones tipo 4, los visitantes que tengan un promedio superior a 80.5 \$ dólares tendrán una permanencia de 137.66 minutos (X4) variable definida en nuestro set de datos). Por otro lado, los visitantes que estén en un promedio inferior a los 137 minutos no serán tomados en cuenta con 115 casos. También se establece que deben tener como mínimo un scrolling de 32.14 (X5) para que sea un visitante de tipo 4, es decir que un visitante de tipo cuatro debe cumplir

con todas estas acciones en un sitio web para poder maximizar la probabilidad de compra con una certeza de que estamos entregando la promoción adecuada del 99.13%.

#### 5.4. PROCESO ADMINISTRATIVO 5

El siguiente proceso que se refiere a la disposición de recursos asignados a las promociones y del tipo de campaña establecidas a visitantes específicos son de hecho, el fin último del presente trabajo como propósito intrínseco de toda actividad empresarial, y encontrar herramientas de gestión adecuada, entonces el último proceso tiene como principal objetivo de tamizar la información vertida por el proceso operativo 4 donde se define los tipos de clientes y la eficiencia del modelo para tomar decisiones. Con el ejemplo anterior vamos a desarrollar los componentes idóneos para establecer los parámetros de modificación y asignación de recursos en campañas digitales.

Figura 27. Proceso administrativo 5



Fuente: Elaboración propia.



Básicamente, el proceso de retroalimentación no está dado por el resultado de las campañas, sino que estarán en función de los parámetros del modelo y su idoneidad para aportar información valiosa en la asignación de recursos y creación de campañas de marketing digital. Por lo cual se establece que el grado de pureza de cada nodo, que en este caso son 6 será sujeto a constante evaluación y dimensionamiento de sus correspondientes etiquetas, para lograr así un adecuado funcionamiento y empleo de recurso tanto de capital humano como de recursos asignados a las campañas de marketing digital.



## CAPITULO VI

### CONCLUSIONES Y RECOMENDACIONES

A partir de las tareas de consecución de la ejecución de un modelo de inteligencia artificial y su incidencia en la identificación de perfiles de suscripción en la educación virtual para la definición de estrategias de asignación de recursos en el marketing digital se determina las siguientes conclusiones y recomendaciones:

#### 6.1. Conclusiones

- El empleo del Machine learning que es un campo de la computación y una rama de la inteligencia artificial, aplicada la identificación de perfiles de suscripción produce un alto impacto en la automatización de información para la creación de estrategias de asignación de recursos publicitarios, debido a que en un promedio del 12% incrementa el retorno de la publicidad (ROI) en esta área de una organización que comercializa cursos virtuales y reduce en un promedio de 5% el costo publicitario total en campañas de marketing digital, tomando como base las redes sociales. Además de otorgar una confiabilidad del 99,13% estableciendo así un parámetro de control de calidad en cuanto al nivel de predicciones y la calidad del dataframe a usarse en el empleo de modelos predictivos en la organización.
- Al aplicar el modelo de machine learning en específico el modelo de clasificación (árboles de decisión) con el gini como principal estándar de calidad y mejoramiento – 0.747 en este caso - se puede concluir que establecer estos parámetros de optimización de dicho modelo deben basarse en la periodicidad de los datos utilizados, es decir, que los datos de entrada deben estar sujetos a comportamiento latentes y observables de los visitantes de un sitio web de educación virtual.
- El desarrollar y establecer el algoritmo de machine learning en un sitio web de educación virtual es primordial evaluarlo en función de su nivel de predicción que en este caso es de 99.13%, este debería enmarcarse en el incremento del ROI (17,85%) y la reducción de los costos publicitarios



(5,85%) asignados en las campañas de marketing digital principalmente en las redes sociales.

- La aplicación de un algoritmo de inteligencia artificial debe basarse en esencia en el tipo de perfil de suscriptor, es decir, en el perfil que visita un sitio web de educación virtual generando así, la construcción adecuada del data set y posteriormente el desarrollo del data frame el cual se emplea en el modelado y ejecución de un algoritmo de machine learning, esto a su vez desarrolla conocimiento y experiencia empresarial para la adopción adecuada de este tipo herramientas en la gestión del marketing repercutiendo en un incremento de la inversión publicitaria y conocimiento de los perfiles de suscripción (buyer persona).
- En relación con la selección del algoritmo de Árboles de Decisión como modelo final ajustado con la totalidad de los datos, se consideraron dos factores, el primero su excelente desempeño con el conjunto de datos de pruebas acorde a su práctica legibilidad referida a la precisión de predicción y el segundo que su costo computacional (trabajo en la nube) además del tiempo de entrenamiento del modelo es relativamente bajo, estas condiciones favorecen la actualización continua del modelo con los nuevos datos adquiridos por el sistema de monitoreo (web scraping) de acciones de un visitante en un sitio web de educación virtual.
- En consecuencia, el objetivo general del presente trabajo de investigación el establecimiento de un modelo de machine learning basado en la estructura de árboles de decisión para la identificación de perfiles de suscripción en la educación virtual la cual permite establecer estrategias de asignación de recursos publicitarios se ha cumplido de manera satisfactoria, debido a que permite identificar los suscriptores en función de su clasificación y así poder desarrollar una estrategia de asignación de recursos publicitarios empleados en las campañas de marketing digital, teniendo como una base la utilización del lenguaje de programación Python incidiendo así en la creación de cursos virtuales y a su vez en el proceso de adopción de esta rama de las ciencias de la computación en la gerencia de plataformas de educación virtual.



- La hipótesis que describe que un algoritmo de Machine learning basado en arboles de decisión genera la identificación de perfiles de suscripción en la educación virtual permitiendo definir una estrategia de asignación de recursos publicitarios enfocados a las campañas de marketing digital es válida en relación al impacto que tiene en el ROI (con un 17,85%) y la reducción del costo promedio del 5,85% asignado en las campañas de marketing digital enfocado a redes sociales y Google Ads, obteniendo así un impacto integral en las áreas de planificación del marketing, estrategias de marketing digital, modelos de gerencia integrados con inteligencia artificial y generación de parámetros en la creación e identificación de perfiles de suscripción en la educación virtual.





## 6.2. Recomendaciones

Lo anterior justifica o permite concluir en la necesidad de implementar el proceso de digitalización y automatización de toda la información que se vierte en un sitio web de educación virtual para la posterior toma de dediciones enfocadas a mejorar los resultados de las campañas de marketing digital; realizado en el marco de la aplicabilidad de la inteligencia artificial presente en la gestión empresarial, más concretamente en el marketing operativo. A partir de la cual parten las siguientes recomendaciones:

- Enfocar la atención en el tiempo de duración de una sesión del visitante en el sitio web, debido a que este hace un uso frecuente de la plataforma de educación virtual generando así familiaridad, encontrando un punto de relación o puente de comunicación directa con el usuario encontrando y desarrollando campañas de publicidad personalizadas con las características y necesidades de cada usuario enfatizando en promociones personalizadas acorde a poder crear su pack de cursos virtuales.
- Planificar la implementación de un modelo de inteligencia artificial para una organización que persigue fines de lucro debe tener como prioridad la modalidad de desarrollo ágil de software basado en el método iterativo e incremental generando así, un proceso enfocado a establecer una ventaja competitiva y el posicionamiento de una plataforma de educación virtual la cual empleará esta herramienta como un generador de información valiosa para la determinación de estrategias gerenciales y de marketing.
- Procesar los lineamientos para la digitalización y posterior adopción de modelos de predicción tanto en inteligencia artificial, aprendizaje supervisado o no supervisado, y automatización de informes, este último como un impulsor de mejora continua en la determinación de nuevos cursos virtuales y detección de comportamientos latentes de los visitantes en un sitio web de educación virtual.



- Por último, definir la utilización de los distintos modelos de machine learning como el de regresión logística y Random Forest que son necesarios para poder encontrar información de vital importancia acorde a los tipos de datos generados por los visitantes de un sitio web de educación virtual para la correcta toma de decisiones tanto en el ámbito de la publicidad, desarrollo de cursos y de innovación en la educación virtual teniendo como apoyo al lenguaje de programación Python y el Cloud Computing permitiendo reducir costos y tiempos de desarrollo de los diferentes algoritmos de inteligencia artificial.



## REFERENCIAS

- DyckhoffHarald. “A Typology of cutting and packing Problems” (1990), European Journal of operation research
- CERT-UK. (2015). Introduction to Social Engineering.
- Lomba Paul “N.P. Linear Programing: An introductory Analysis” , Ney York, 1964.
- Chang, M. (2017, noviembre 30). 4 Stages of the Machine Learning (ML) Modeling Cycle
- Steiner Rudolf “Análisis de Sistemas”, 1994.
- A Comprehensive Guide to Machine Learning Soroush Nasiriany, Garrett Thomas, William Wang, Alex Yang Department of Electrical Engineering and Computer Sciences University of California, Berkeley June 24, 2019
- FILIBA, Salvador. PALMERI, Ricardo. Manual de Marketing directo e interactivo. 3ra Edición: Asociación de Marketing Directo e Interactivo de Argentina, AMDIA, Buenos Aires. 2008
- Feature Engineering for Machine Learning, Alice Zheng and Amanda Casari 2018.
- PORTER, Michael. Estrategia Competitiva. 1ra Edición: The Free Press. 1980
- LAMBIN, Jean-Jaques. Marketing Estratégico. 3ra Edición: Editorial Mc Graw Hill, Madrid, 1993
- SANTANDREU, Eliseu. Finanzas para directivos de marketing. 1ra Edición: Ediciones Gestión 2000, Barcelona, 2003
- LENSOLD, James D. Marketing ROI. 1ra Edición: Mc Graw Hill, México, 2004
- GIRARD, Bernard. El modelo de Google. 1ra Edición: Editorial Norma, 2009.
- WARREN, Jaques. ROI in online marketing and web analytics. White Paper: AT Internet Online Intelligence Solutions, 2012



# ANEXOS