TRABAJO DE GRADO
Opción Seminario-Diplomado.

Proyecto Aplicación Machine Learning En La Agencia De Turismo Agentur

Corporación Universitaria Remington.
Facultad de Ingeniería
Seminario Machine Learning

Estudiante Juan David Zapata Gallego
Estudiante Julián Andrés González Vásquez
Tutor Juan Pablo Vélez Uribe
Seminario
2023

Dedicatoria

A nuestras familias
A los tutores
En especial a nuestras esposas e hijos

Agradecimientos

A nuestras familias por ser ese motor que nos impulsa a seguir adelante a pesar de las dificultades y del sacrificio de tiempo que esto significa, donde se deja de compartir o dedicar tiempo a las familias, para sumergirnos de lleno en nuestro trabajo, estudio y por consiguiente este proyecto, tratando de tener siempre la mejor actitud para que después se vea reflejado todo ese esfuerzo en una mejor calidad de vida del grupo familiar.

A los tutores por el acompañamiento brindado durante todo este proceso de aprendizaje, que, a pesar de haber tenido vario percances, se logró sacar adelante de la mejor manera, gracias al compromiso de nuestro tutor Juan Pablo Vélez Uribe que siempre tuvo la voluntad y conocimiento para despejar dudas que surgían durante el desarrollo del proyecto con la mejor actitud.

A nuestras esposas que siempre nos apoyaron durante este proceso de aprendizaje, motivándonos siempre a continuar con la mejor actitud y amor incondicional.

Tabla de Contenidos

Resumen

Para llevar a cabo este proyecto, realizamos una investigación exhaustiva sobre los modelos más comunes para el análisis de datos en Machine Learning, centrándonos en el seminario proporcionado en la plataforma Crehana. Esta fuente ha sido clave para obtener la mayor parte de la información. Nuestro enfoque se dirige hacia la aplicación práctica de estos conocimientos en el sector turístico, específicamente en la agencia de viajes "Agentur". A pesar de los desafíos que enfrentó el sector turístico durante la pandemia, actualmente es una de las industrias que experimenta un notable crecimiento debido a una alta demanda. Este crecimiento ha impulsado la proliferación de agencias, principalmente virtuales, con plataformas avanzadas de autogestión que ganan protagonismo frente a las agencias físicas. Motivados por esta tendencia, decidimos implementar un proyecto de Machine Learning que permita a la agencia "Agentur" implementar herramientas tecnológicas para estar a la altura de las agencias virtuales. Comenzamos recopilando datos e información, posteriormente filtrándose y procesándolos con el objetivo de mejorar significativamente la experiencia de los clientes de la agencia Agentur.

Se aborda los fundamentos del Machine Learning, destacando el uso de Python y herramientas como Numpy, Pandas, Scikits Learn, y Jupyter Notebooks. Se detalla el proceso para aplicar Python en ejercicios de regresión, incluyendo la descarga de librerías, asignación de nombres a variables y la transformación de variables no numéricas a Dummy. Se explica el concepto de aprendizaje supervisado, sus categorías y variables dependientes e independientes. Se mencionan algoritmos como regresión lineal, regresión logística, árboles de decisión, bosque aleatorio y redes neuronales. Se introduce el concepto de contrafactual y efecto causal, señalando la limitación del Machine Learning en encontrar relaciones causales. Se exploran algoritmos causales como Double LASSO, Causal Trees y Causal Forest. El texto también aborda la evaluación de modelos, el workflow de Machine Learning, la preparación de datos, la etapa de entrenamiento y métricas de desempeño como la matriz de confusión. Se presentan métodos de validación cruzada como Holdout Cross Validation.Se discuten principios del Machine Learning como la generalización, la navaja de Ockham y conocimiento jerárquico. Se explora la evolución del lenguaje natural, destacando GPT-3 y Github Copilot como herramientas potentes. Se menciona la importancia de la innovación tecnológica con inteligencia artificial, los algoritmos evolutivos y la lógica difusa.

Este proyecto revela la convergencia estratégica entre la innovación tecnológica y la industria turística. Con el creciente papel de la inteligencia artificial, y más específicamente del Machine Learning, en la toma de decisiones empresariales, este proyecto busca transformar la forma en que la agencia comprende y se relaciona con sus clientes. La esencia del proyecto se centra en la anticipación de las necesidades y deseos del cliente a través de modelos predictivos de Machine Learning. Al abordar la capacidad de prever el interés de los clientes en realizar segundas y terceras compras, así como en la evaluación de patrones de interés, preferencias de destinos y estrategias de marketing, se plantea una visión innovadora para la mejora de la fidelización del cliente. Este enfoque no solo busca aumentar las ventas y optimizar las campañas de marketing, sino que también pretende proporcionar una experiencia más personalizada y satisfactoria para cada cliente. Al alinear la tecnología con las expectativas del cliente en la industria de

viajes, se aspira no solo a la eficiencia operativa, sino a la creación de relaciones más sólidas y duraderas, llevando la fidelización del cliente a un nuevo nivel.

Palabras clave

Machine Learning, Big Data, Inversión tecnológica, aprendizaje automático, regresión logística, aprendizaje supervisado, la ética en la inteligencia artificial, análisis contrafactual, clustering, redes neuronales.

Marco conceptual y contextual

Machine Learning

Conceptos introductorios.

El desarrollo de este proyecto final está orientado al estudio de Machine Learning y su aplicación en la industria turística, específicamente en las agencias de viajes y en especial a la agencia llamada viajes agentur. Machine Learning es un campo dentro de la inteligencia artificial que reúne algoritmos capaces de aprender e imitar un comportamiento racional, acorde a una situación de manera autónoma. Su principal característica es la capacidad de mejorar automáticamente sus comportamientos basados en la experiencia y sin la intervención humana. [1]

El Machine Learning es una herramienta muy poderosa sobre todo en la época en la que vivimos ya que gracias a internet tenemos a nuestro alcance una gran cantidad de datos e información y el Machine Learning tiene el poder de ordenar y organizar esta información para que podamos hacer uso de ella volviéndose el asistente personal de cada persona, el cual aprende a conocer todos nuestros gustos con base en nuestra navegación en la web y en la utilización de los diferentes dispositivos y plataformas vinculadas a la red, algunas de las aplicaciones utilizadas frecuentemente son Netflix, YouTube, Amazon, la banca electrónica, el correo electrónico, motores de búsqueda como Google y aplicaciones como Facebook, entre muchas otras que frecuentemente usamos.

El Machine Learning se basa en los pronósticos, para ello se basa en los datos "Data Scientist" el cual estará encargado de las técnicas para el uso, el análisis y la presentación de los datos de forma tal que se pueda llegar a predicciones más acertadas.

El aprendizaje automático es muy útil para filtrar correos no deseados al igual que filtra otros que son potencialmente peligrosos, también es de gran ayuda en el texto predictivo para autocompletar palabras y ayudarnos a escribir mucho más rápido, con lo cual el ahorro de tiempo es un valor agregado enorme para las funciones que requieren de estar constantemente escribiendo mensajes.

Si bien el aprendizaje automático o Machine Learning puede ser beneficioso para el ocio y la seguridad de los datos, presenta algunos casos de uso poco ético.

Lo que nos muestra que esta tecnología también tiene potencial para ser utilizada de manera cuestionable por empresas y gobiernos, para predecir el comportamiento, rastrear y reprimir personas, por lo que es necesario prohibir el uso de esta tecnología en aplicaciones como las antes mencionadas y regular su uso en las demás aplicaciones, para evitar situaciones similares en el futuro, ya que las repercusiones podrían ser desastrosas sí se utiliza para actividades ilícitas y es más el daño que puede ocasionar que los beneficios que tiene en la actualidad, esto puede ocurrir igual con cualquier herramienta o tecnología humana. [2]

El término Big Data hace referencia a la acumulación masiva de datos, llevado al punto de superar la capacidad de las herramientas tradicionales para ser registrados, gestionados y procesados en un tiempo razonable. Por lo tanto, un conjunto de datos se le categoriza dentro de Big Data cuando es demasiado grande para ser manejado de forma apropiada por los programas convencionales de software. Los datos a gran volumen deben tener capacidad de volumen, velocidad, el valor de la información, su veracidad y es el insumo principal para la aplicación del Machine Learning.

Ya teniendo claros los conceptos anteriores, es importante destacar que Machine Learning depende en su totalidad del uso los datos y por eso la importancia de una fuerte inversión en tecnología, dándole prioridad a esta última ya que es la base para poder dar inicio a un proyecto de Machine Learning, sin la tecnología adecuada es casi imposible que un proyecto tenga futuro. [3]

El objetivo principal de Machine Learning es generar formas de traducir características de un grupo de observaciones y transformarla en predicciones que van a ayudar a la toma de decisiones.

Por otro lado, está el turismo que ha venido evolucionando desde principios de siglos fruto del incremento del tiempo libre propiciado por el desarrollo tecnológico en las sociedades de consumo de los países desarrollados.
La consolidación de la sociedad industrial, la mecanización y la robotización progresivas permiten la ampliación del tiempo de ocio al conjunto de la población y es allí donde se incorpora el ocio a una nueva escala de valores, dando mucha más relevancia en el mercado turístico.

Viajes agentur es una compañía de la industria turística con 75 años de experiencia en el mercado, actualmente es socia de L'ALIANXA TRAVEL NETWORK y GLOBAL STAR que permiten tener presencia en más de 90 países y contar con 3000 oficinas para garantizar un servicio de cobertura global de alto nivel, es allí donde nace la necesidad de incorporar nuevas tecnologías que permitan su permanencia en el tiempo como la hecho durante estos 75 años de servicio a las empresas y comunidad que frecuentemente requieren de su asesoría para materializar sus sueños y necesidades de recorrer cada rincón del mundo con un acompañamiento constante.

Para aplicar Machine Learning se debe tener claro cuál es el objetivo que se quiere lograr, que se busca resolver, investigar si es una aplicación viable económicamente, esto dado a que el 87% de los proyectos de Machine Learning no ven la luz del día dado a que no generan valor.

Para analizar si es viable o no su aplicación debemos utilizar la pirámide de valor. Como primer paso en la base de la pirámide tenemos la tecnología que le permitirá a la agencia de viajes "Agentur", procesar, mantener y guardar los datos, la agencia debe contar con la tecnología capaz de correr modelos de Machine Learning, luego seguimos con la gobernanza de los datos, donde se necesita que la agencia cuente con personas especializadas en extraer, transformar y almacenar los datos constantemente para que la

información esté disponible en el momento que se necesite estos deben tener calidad, un linaje y que se almacenen de forma segura. Continuando en la pirámide tenemos las operaciones, luego la inteligencia de los negocios donde encontramos tableros útiles que nos presentan la información de una forma intuitiva como lo es Power BI y en el tope de la pirámide encontramos la inteligencia artificial como motor para tomar decisiones o que nos ejecute predicciones. [4]

Un aspecto importante a tener en cuenta cuando se quiere implementar Machine Learning es encontrar las personas idóneas, en este campo existe una alta rotación del personal debido a las expectativas versus la realidad, esto se puede deber a aspiraciones de los estudiantes de las áreas de tecnología donde pretenden generar algoritmos los cuales ya pueden existir, en el mundo laboral es normal que se encuentre con funciones como preparar, limpiar y procesar la data que permite generar valor a la información. [5]

El valor cuando se adquiere conocimientos en tecnología no es la generación de los algoritmos; sino en la cantidad, calidad, limpieza y depuración de los datos para que los modelos y los algoritmos funciones de forma correcta.

La correcta implementación del Machine Learning una vez se decide aplicarla y que se cuente con la persona adecuada no es tratarla como un área aparte, esta debe estar articulada a todos los procesos de la empresa donde se involucre la información, por tal motivo es importante que las personas que lo vayan a implementar sean propias de la empresa de turismo y recurrir a consultores dado el caso. [6]

Todos los roles a la hora participar en la construcción de un modelo de Machine Learning es importante, las habilidades en programación son necesarias, pero también lo son el conocimiento del negocio, en este caso el del turismo. Debe existir un trabajo en equipo donde se brinde la experiencia del ingeniero de Machine Learning como también la experiencia en el campo del turismo; es aquí donde generalmente entra una figura que se llama el Product Owner, quien levanta la información con las personas expertas y luego la lleva a los científicos de datos para la creación de los modelos por medio de las historias de usuario.

Como se ha mencionado anteriormente, antes de aplicar la herramienta debemos tener claro cuál va a ser nuestro objetivo por predecir, los datos deben estar divididos en observaciones y número de variables que servirán para responder a nuestra variable objetivo, con el tiempo el número de variables y observaciones seguirán creciendo lo que permitirá que nuestra variable objetivo sea más exacta en la predicción positiva que queremos.

En Machine Learning existen varios tipos de algoritmos; el de clasificación, el de regresión , el de clusterización y los de asociación. El primero consiste en aquellas clasificaciones dónde las observaciones son A, B, C. Etc. Lo que lo hace binaria respondiendo a si o no, el segundo está enfocado a variables continuas como por ejemplo la predicción del tiempo, el tercero busca encontrar estructuras inherentemente presentes

en los datos y organizarlos de acuerdo con la similitud y el de asociación se utiliza para descubrir patrones de relación entre diferentes elementos en un conjunto de datos. [7]

Un algoritmo de regresión lineal permite generar predicciones en base a promedios, sin embargo, este tipo de aplicaciones pueden no resultar efectivas cuando el volumen de la información es considerable. [8]

Una herramienta que utilizamos y que es fácil su utilización es Excel, sin embargo, como se mencionó anteriormente cuando el volumen es muy grande es necesario recurrir a lenguaje SQL donde a través de QUERY se solicita información y este se responde por medio de bases de datos que pueden resultar de varias fuentes.

Este tipo de bases de datos sean manejados en Excel o en SQL nos brinda información estructurada y de un manejo más sencillo, sin embargo, también podemos tener datos no estructurados los cuales constan de audios o imágenes y que la regresión lineal no le será posible predecir dada a su complejidad.[9]

Fundamentos del Machine Learning.

Para la aplicación de Machine Learning existen diversos Software, entre los más populares tenemos R y Python.

La más utilizada es Python, para su correcta aplicación debemos contar con algunas herramientas que son útiles para correr los algoritmos, entre ellas tenemos:

Numpy: Encargada de habilitar la computación numérica en Python.

Pandas: Analiza los datos de una forma práctica para el mundo real.

Scikits Learn: Analiza los datos predictivos y se integra a paquetes de Numpy y Pandas.

Jupiter Notebooks: Es la interfaz que se instala en nuestro navegador, permite la mezcla de textos, ecuaciones y gráficos y sirve para documentar el proceso. [10]

Para la aplicación de Python en ejercicios de regresión se debe seguir una serie de procedimientos vistos en el seminario los cuales nos permiten desarrollar el modelo de Machine Learning. Entre estos procedimientos y herramientas aprendidos tenemos:

- Descargar la librería y entender la data.
- Asignar el nombre a las variables, estas deben ser entendibles.
- Hallar la variable objetivo en base a las variables predictivas.
- Las variables categóricas, ósea, que no son numéricas deben ser cambiadas a una variable Dummy, al ejecutarlo en la programación nos permite convertir esta variable no numérica a asignarle un número.
- Analizar qué variable se puede predecir la función objetivo, entre menos variables será más sencillo aplicar el modelamiento y luego ir. A medida que el

modelamiento sea estable se puede incluir más variables que permita incrementar el nivel de pronósticos.

- En caso de tener cálculos estadísticos, la herramienta Numpy nos puede servir. Las gráficas nos pueden mostrar los valores mínimos, máximos y la media.
- Aplicación de la matriz de correlación: esta permite ver todas las variables en una sola gráfica las cuales van entre -1 a 1. 0 indica que no tiene relación. -1 indica que la otra aumenta a 1 mientras que 1 indica que la otra disminuye a -1. En la figura 1 podemos observar como ejemplo su composición.

Figura 1: Ejemplo matriz de correlación en Python. [11]

- Para correr el modelo se recomienda un 80% para entrenamiento y un 20% para realizar test o pruebas finales, luego la pasamos a la métrica de regresión, entre ellas tenemos RMSE que tiene presente las diferencia entre los valores reales y las predicciones para cada una de las observaciones. y R2 que explica que variaciones hay con la variable objetivo. [12]

Metodología para la aplicación de algoritmos de clasificación en la empresa de viajes "Agentur"::

- Crear la Data Frame con variables que sean intuitivas.
- Las variables explicativas permiten encontrar la variable objetivo, para ello es necesario conocer qué problema es el que se pretende resolver o por qué se va a aplicar.
- Si se tienen variables categóricas se debe convertir en variables numéricas, para ellos se puede utilizar un Label Esconder para cada una de las variables.
- Utilizar la matriz de confusión: Esta matriz sirve para algoritmos de clasificación y muestra el número de clasificaciones correctas a cada categoría.
- utilizar un 80% de entrenamiento y 20% para la evaluación del modelo.

- Una vez aplicada la matriz de confusión es posible predecir la precisión del modelo. [13]

Una vez se termina el modelo más adecuado a aplicar de Machine Learning debemos tener en cuenta su ciclo de vida:

- Definir el objetivo y el valor que se quiere agregar.
- Recolectar la información, esta requiere esfuerzos que permitan encontrar los datos necesarios, su procedencia y entenderla al igual que filtrar los puntos legales.
- Entrenar el modelo, luego de tener la data es necesario encontrar los mejores algoritmos que se adapten, ya sea regresión lineal o logística.
- Producción como parte final del ciclo, estos modelos deben alinearse con otros softwares que permitan visualizar las decisiones que se deben tomar o realizar seguimiento cuando se tratan de modelos de regresión.
- Evaluación y refinamiento del modelo [14]

Un aspecto a tener en cuenta para la aplicación del modelo es si se trata de un aprendizaje supervisado, el cual se alimenta de datos estructurados como los que encontramos en tablas con filas y columnas. Cuando se trata de un aprendizaje no supervisado o clustering son datos que se clasifican por semejanzas, estas pueden segmentar los clientes y en base a ello tomar decisiones estratégicas de marketing que se basa en el patrón de comportamiento que tenga cada cliente.

Conceptos básicos de aprendizaje supervisado

¿Qué es el aprendizaje supervisado?

El aprendizaje supervisado, también conocido como Machine Learning supervisado, es una subcategoría de Machine Learning y la inteligencia artificial. Se define por su uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifican datos o prevén resultados con precisión. A medida que se introducen datos en el modelo, ajusta sus ponderaciones hasta que el modelo se adapte correctamente, lo que ocurre como parte del proceso de validación cruzada. El aprendizaje supervisado permite a las organizaciones resolver una amplia variedad de problemas del mundo real a escala como, por ejemplo, la clasificación de spam en una carpeta distinta de la bandeja de entrada**.** [15]

Por medio de este aprendizaje supervisado podemos desarrollar modelos de predicciones dobles, para el caso del proyecto se puede realizar una primera predicción y luego una

clasificación segmentada la cual puede variar de acuerdo con el comportamiento que tenga el cliente a través del tiempo.

El Machine Learning se divide en varias categorías en la cual encontramos el aprendizaje supervisado la cual se basa en un agrupamiento de los datos, el aprendizaje no supervisado que se basa en un patrón de datos y el aprendizaje semi supervisado donde encontramos las redes neuronales.

En el aprendizaje supervisado encontraremos variables dependientes y no dependientes, Los datos son importantes para desarrollar el modelo, lo más comunes son las tablas de datos transaccionales.

Cómo funciona el aprendizaje supervisado

El aprendizaje supervisado utiliza un conjunto de datos de entrenamiento para enseñar a los modelos a generar la salida deseada. Este conjunto de datos de entrenamiento incluye entradas y salidas correctas que, a su vez, permiten que el modelo aprenda con el tiempo. El algoritmo mide su exactitud a través de la función de pérdida, ajustándose hasta que el error se ha minimizado lo suficiente.

Algoritmos de aprendizaje supervisado

En los procesos de Machine Learning supervisado se utilizan varios algoritmos y técnicas de cálculo. A continuación, vamos a incluir unas explicaciones sobre los métodos de aprendizaje más utilizados bajo la herramienta Python. También nos permitirá decidir cuál es el algoritmo más adecuado para aplicar al proyecto.

Es importante determinar qué tipo de variable tenemos y así aplicar las métricas de desempeño las cuales nos brindaran el porcentaje de exactitud para cada uno de los algoritmos que aplicaremos.

En las variables continuas encontraremos métricas de desempeño como son MAE, MSE y RMSE. En las variables categóricas, aplicaremos las métricas ACCURACY, F1 y CURVAS DE ROC.

Cuando corremos el algoritmo nos brindara un valor predicho y el valor real, adicional a las métricas antes mencionadas por cada tipo de variable, podemos aplicar la métrica de asertividad que nos dice que tanto de nuestros datos se clasificaron correctamente versus los que no. La matriz de confusión por medio de la herramienta CF nos brinda gráficamente los valores que están clasificados y los que no y por último la métrica de precisión nos dice los valores que se clasificaron versus los que no, más los que se debieron clasificar en otra categoría. [16],[17]

La herramienta “train test split” permite dividir los datos de entrenamiento y los datos de prueba, también nos mostrará como x=variable independiente, y=variable dependiente y

test size la cantidad de datos a utilizar como se ha mencionado anteriormente, se recomienda un 30%. [18]

Entre los algoritmos aplicados tenemos:

Regresión lineal: la regresión lineal se utiliza para identificar la relación entre una variable dependiente y una o más variables independientes y, por lo general, se aprovecha para realizar predicciones sobre resultados futuros. Cuando solo hay una variable independiente y una variable dependiente, se conoce como regresión lineal simple. A medida que aumenta el número de variables independientes, se habla de regresión lineal múltiple. Para cada tipo de regresión lineal, trata de trazar la línea que mejor se ajuste, la cual se calcula mediante el método de mínimos cuadrados. Sin embargo, a diferencia de otros modelos de regresión, esta línea es recta cuando se traza en un grafo.

Para que las variables independientes estén lo más ordenadas posible podemos utilizar el algoritmo de mínimo cuadrados ordinarios el cual ajusta los datos. [19]

Regresión logística: mientras que la regresión lineal se utiliza cuando las variables dependientes son continuas, la regresión logística se selecciona cuando la variable dependiente es categórica; esto quiere decir que tiene salidas binarias, como "verdadero" y "falso" o "sí" o "no". Ambos modelos de regresión tratan de comprender las relaciones entre las entradas de datos, pero la regresión logística se utiliza principalmente para resolver problemas de clasificación binaria, como la identificación de correo no deseado.

Dado a que la aplicación de la regresión logística es para variables de clasificación es importante tener en cuenta que las palabras o textos deben ser convertidos a números. [20]

Árboles de decisión: Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquico, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Este algoritmo permite clasificar segmentos de acuerdo con los datos. [21]

Bosque aleatorio: el bosque aleatorio es otro algoritmo flexible de Machine Learning supervisado que se utiliza tanto para la clasificación como para la regresión. El "bosque" hace referencia a una colección de árboles de decisiones no correlacionados, que se fusionan para reducir la varianza y crear predicciones de datos más precisas.

También podemos aplicar el random forest-relevancia de variables para conocer el peso de un conjunto de variables y así determinar cuáles son las más adecuadas para resolver nuestra variable objetivo.[22]

Redes neuronales: las redes neuronales, que se utilizan principalmente para los algoritmos de Deep Learning, procesan los datos de entrenamiento imitando la interconectividad del cerebro humano a través de capas de nodos. Cada nodo está formado por entradas, ponderaciones, un sesgo (o umbral) y una salida. Si ese valor de salida excede un umbral determinado, "dispara" o activa el nodo, pasando datos a la

siguiente capa de la red. Las redes neuronales aprenden esta función de correlación a través del aprendizaje supervisado, y se ajustan con base en la función de pérdida a través del proceso de su gradiente descendente. Cuando la función de coste es igual o se acerca a cero, podemos confiar en la precisión del modelo para obtener la respuesta correcta.

En las redes neuronales constan de tres partes, un input layer que permitirá su entrenamiento, un hidden layer que es la parte de la activación y por último un output layer que nos brindara la predicción o clasificación.

Cuando aplicamos este algoritmo para procesos de clasificación podemos encontrar que existen varios output layers, para ello se recomienda utilizar la herramienta to:categorical la cual permitirá asignar varias salidas de acuerdo con su clasificación. [23]

Introducción al análisis causal

¿Qué es un contrafactual?

Contrafactual en términos más sencillos se puede describir como el "hubiera", que es en sí, los resultados de una acción que no sucedió, esto suele pasar muy a menudo cuando se tienen diferentes alternativas y finalmente se elige una de todas y queda la pregunta. ¿Qué hubiera pasado si de las alternativas, se hubiera seleccionado otra diferente? Todos estos cuestionamientos dan lugar a un contrafactual. En el caso que se pudiera saber qué hubiera pasado en un escenario alternativo se podrían estimar resultados potenciales y a esto se le llama el efecto causal.

El efecto causal es una comparación entre dos estados: uno que realmente sucedió y otro imaginario donde no sucedió. El papel que juega el contrafactual es muy importante porque permite establecer relaciones causales entre acciones y los resultados de dichas acciones, debido a que mide el efecto entre dos variables y las compara cuando una de las dos se encuentre ausente.

Es importante destacar que el Machine Learning es muy poderoso para encontrar relaciones en los datos, pero no relaciones causales, lo cual no hace del todo confiable el Machine Learning para la toma de decisiones de negocio ya que por sí solo no permite saber la influencia que tiene una variable sobre la otra. [24]

En el análisis contrafactual es importante conocer el significado de algunos términos como la correlación, que hace referencia a la presencia o ausencia de cualquier relación lineal entre dos variables: Otro término es la causalidad que hace referencia a causa efecto, donde un evento está causado por uno anterior. Es importante aclarar que entre estos dos términos la correlación implica asociación, pero no causalidad y la causalidad implica asociación, pero no correlación. Otro término es la variable omitida, que puede ser observable y no observable. [25]

Variable omitida observable: Es una variable que no se consideró al momento de realizar el análisis

Variable omitida inobservable: Es la variable que no pudo ser observada en los datos

Otro término es la variable omitida, que puede ser observable y no observable.

Es importante tener en cuenta los sesgos por variables omitidas, entre ellas las observables, las cuales no se tuvieron en cuenta a la hora de realizar el análisis y las inobservables que no se pueden ver en nuestros datos. Son situaciones que debemos tener en cuenta en la implementación del proyecto. Un caso que se puede presentar es a la hora de analizar nuestros datos de los clientes con sus respectivas variables explicativas y no tener en cuenta habilidades que pueda tener el cliente y que puedan servir a la hora de generar las predicciones, como por ejemplo si se le va a ofrecer paquetes turísticos donde se incluyan deportes. [26]

Churn: hace referencia a la tasa de abandono de productos que los usuarios dejan de utilizar en determinado periodo de tiempo. Es muy importante saber la tasa de abandono porque los costos de adquisición de un cliente son más altos que los costos de retención, por los que es más viable crear estrategias para prevenir el abandono.

Sesgo de selección: El sesgo de selección hace referencia a los errores experimentales que conducen a una representación inexacta de la muestra de la investigación. Surge cuando el grupo de participantes o los datos no representan al grupo objetivo.

Para ello es importante que el grupo esté estratificado en varias variables, no demasiadas, pero sí las suficientes para evitar que se genere una asignación sesgada.

Una causa importante del sesgo de selección es cuando no se tiene en cuenta las características de los subgrupos. Provoca disparidades fundamentales entre las variables de los datos de la muestra y la población de la investigación. [27]

Aprendizaje de máquina

Dando un repaso a lo que es Machine Learning o aprendizaje de máquina, este nos permite generar predicciones por medio de herramientas y se alimenta de una base de datos con unas variables de intereses. Su poder predictivo puede mejorar a través de la llegada de nuestros datos.

Como se ha mencionado anteriormente, la base de datos debe ser dividida en dos para evitar el error cuadrático medio fuera de la muestra, para ello tenemos un porcentaje de la base de datos que se utilizarán para entrenar el algoritmo, y el otro porcentaje para validar la calidad de las predicciones. Cuando se utiliza gran porcentaje de los datos para

entrenar el algoritmo puede ocasionar que las predicciones no sean exactas cuando se incluyen nuevos datos. [28]

La inferencia causal cuando tenemos Big Data nos permite obtener clasificaciones insesgadas dado a que la gran cantidad de datos puede tener en gran cantidad de variables que quizás no teníamos en cuenta a la hora de generar el modelo. Con esta ventaja podemos anticiparnos a las reacciones de los clientes frente al ofrecimiento de los paquetes turísticos y brindar asesorías más personalizadas. [29]

Algoritmos causales:

Double LASSO: Permite seleccionar las variables más precisas que nos explique la variable de interés, para ello se selecciona las que están más correlacionadas, luego se estima una regresión por el método de mínimos cuadrados. Esto lo que permite es minimizar el sesgo por variables omitidas y el estimador se vuelve más preciso. [30]

La aplicación del algoritmo nos va a permitir hallar soluciones ante situaciones complejas donde no es fácil determinar que tanto un cliente va a comprar un paquete turístico teniendo en cuenta sólo la variable de capacidad económica, por ejemplo. Sino que se tendrá en cuenta otro tipo de variables como la atención durante las asesorías y que puede influir en la compra.

Causal Trees: Por medio de árboles de decisión en el nodo se crean grupos de personas con características similares, uno de los grupos es tratada y el otro no, luego se estima la diferencia entre el grupo de tratamiento y el grupo de control. [31]

Causal Forest: Basado en los Causal Trees, su funcionamiento es encontrar efectos heterogéneos, a su vez, puede mejorar los resultados del Double Debiased ML cuando hay un impacto no lineal. Como desventaja al ser un algoritmo complejo puede presentar un alto consumo de recursos computacionales. [32]

Evaluación de modelos de Machine Learning

Workflow: Este consta de varias etapas, primero se encuentra el entendimiento del negocio, cuál es su objetivo y las necesidades que deben atenderse, luego pasamos al entendimiento de los datos, con cuales contamos y de qué tipo son, una vez entenderlo pasamos a la preparación de los datos, luego pasamos al modelado donde entrenaremos nuestro algoritmo de Machine Learning, una vez modelado lo evaluamos y para finalizar la parte productiva del modelo de Machine Learning. [33]

Preparación de los datos: Para correr correctamente un modelo de Machine Learning es necesario preparar el Dataset, para ello podemos recurrir a herramientas de programación que nos permitirá visualizar los datos, entre ellos el número de observaciones y el número de variables, para datos numéricos la mínima, la máxima, la desviación estándar, la

media. También dar un tratamiento a los valores ausentes y sustituirlos por valores promedios si así lo elegimos o con los valores más frecuentes.

También es necesario convertir las variables categóricas en datos independientes en cuanto a las variables, para ello podemos utilizar herramientas get_dummies, luego para las variables numéricas utilizaremos la herramienta MinMaxScaler con el fin de escalar y homogeneizar las variables numéricas. [34]

Etapa de entrenamiento: Conlleva un 20% del proyecto de Machine Learning, Se importa el dataset, se debe recordar en dividir los datos de entrenamiento y los datos de prueba. Una vez ejecutado el algoritmo procedemos a las predicciones las cuales intenta capturar la relación que se tiene entre las variables predictoras con la variable objetivo. Se recomienda que cuando dos algoritmos tienen un comportamiento similar se escoja el más sencillo de ejecutar. [35]

Métricas de desempeño

Matriz de confusión: Resume el desempeño que tiene el algoritmo en los modelos de clasificación, esta matriz nos permite visualizar el resultado de las predicciones para cada una de las observaciones [36]. Lo que nos interesa en la matriz son los valores dados en la diagonal, es decir los verdaderos positivos y los verdaderos negativos como se observa en la Figura 2.

Figura 2: Matriz de confusión. [37]

Accuracy: Derivada de la matriz de confusión, está resume el número de predicciones correctas dividida el total de predicciones.

Su fórmula es (TP+TN)/TN+FP+FN+TP), el resultado nos brindará el porcentaje de efectividad que se tiene en las predicciones.

Su desventaja es que oculta la clasificación de las observaciones lo que puede ocasionar que genere un porcentaje alto en etiquetas negativas; sin embargo, esto se puede solucionar con otras métricas que se pueden obtener con la matriz de confusión. [38]

Precisión: Esta métrica se calcula dividiendo el número de verdaderos positivos y el número total de verdaderos positivos y falsos positivos. [39]

Recall: es el resultado de dividir el número de verdaderos positivos y la suma de los verdaderos positivos y los falsos negativos. [39]

Specificity: Es el caso contrario al Recall, en este caso su resultado es la división de los verdaderos negativos y la suma de los verdaderos negativos y los falsos positivos. [40]

F1 score: Combina la precisión y el Recall y es el resultado de dividir la precisión por el Recall y la precisión más el Recall, al resultado se le multiplica por 2. [40]

Curva de ROC y AUC: Utilizada en problemas de clasificación, este tiene en cuenta el área bajo la curva, entre mayor sea el área mejor será el comportamiento del modelo. [41]

En la Figura 3 podemos observar la inclinación que tiene la curva superior la cual indica que es buena, a medida que la curva disminuye el comportamiento del modelo disminuye.

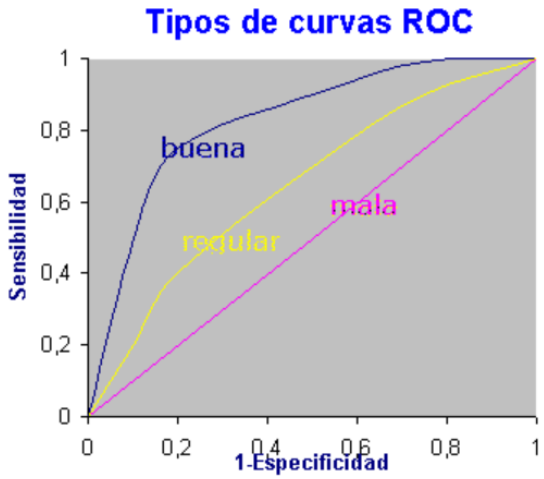

Figura 3: Curva de ROC y AUC. [42]

Curva de Precisión Recall: Permite evaluar el rendimiento para modelos de clasificación binaria donde se grafica en el eje y precisión y Recall en el eje de la x. [43]

Métodos de validación cruzada

Permite realizar un muestreo entre los datos de entrenamiento y los datos de validación que nos permite mirar que tan efectiva es la capacidad de predicción correcta, también

para estimar el comportamiento ante nuevos datos que no se tuvieron en cuenta en el modelo [44]. Entre estos métodos tenemos:

- **Holdout Cross Validation:** Ambos sistemas ayudan a evaluar el rendimiento del modelo de manera más confiable, que solo evaluar y probar con un solo conjunto de datos.
- **K-fold Cross Validation:** Esta técnica evalúa el rendimiento de una manera más robusta.
- **Stratified K-fold Cross Validation:** Se utiliza cuando se quiere asegurar que la proporción de clases en cada fold sea representativa del conjunto de datos original,
- **Leave-p-out Cross Validation:** Esta técnica implica dejar por fuera un conjunto específico de observaciones del conjunto de entrenamiento para evaluar el rendimiento del modelo.
- **Leave-one-out Cross Validation:** En esta técnica se utiliza un solo dato como conjunto de prueba y el resto se utiliza para entrenar el modelo.
- **Rolling Cross Validation:** Es una estrategia en la que se va actualizando continuamente el conjunto de entrenamiento y de prueba en una ventana temporal.
- **Monte Carlo Cross Validation:** Esta es una variante de la validación cruzada en la que se realizan múltiples peticiones de manera aleatoria.

Introducción a la inteligencia artificial.

Principios del Machine Learning

- **La generalización:** Este principio hace referencia a no buscar modelos exactos para que este modelo sea útil en situaciones del mundo real por medio de modelos matemáticos.
- **No free lu:** Lo que se quiere dar a entender con este término es que la inteligencia artificial no está en la capacidad de resolver problemas en su totalidad y se requiere de utilizar el algoritmo adecuado para cada tarea en particular.
- **La navaja de Ockham o principio de parsimonia: La** idea fundamental es que entre más sencillos sean los modelos y algoritmos, es mucho mejor.
- **Más datos es mejor que tener algoritmos complejos:** Tener un conjunto de datos más grande puede ser más beneficioso ya que un aprendizaje automático puede aprender patrones más robustos cuando se entrena con una cantidad significativa de datos, con la implementación de algoritmos muy precisos.
- **Validación cruzada:** Se deben hacer validaciones cruzadas para verificar su funcionamiento de forma generalizada.
- **Diversidad algoritmo:** Es importante explorar varios algoritmos para escoger el mejor de todos y el más indicado. [45]

Conceptos alrededor de la inteligencia artificia

- **IA:** La inteligencia artificial se puede dividir en una general que es la capacidad que tienen las computadoras para imitar las capacidades cognitivas de los seres

humanos, por otro lado, tenemos la inteligencia artificial Narrow que tiene la capacidad de reconocer texto, audio o también imágenes.

- **Machine Learning:** Es un aprendizaje automático que aprende de la misma forma que los seres humanos y es a través de los datos, cuan más repetitivo pueda ser es mejor dado a que puede tener un óptimo aprendizaje del cual permitirá tener la capacidad de hacer predicciones correctamente y como ya se mencionó antes, es un subcampo de la IA.
- **Deep Learning:** Es una capacidad más profundamente matemática, donde aprende no solo de ejemplos sino de representaciones y características complejas. Gracias a la interacción que se tiene constantemente entre las computadoras y los seres humanos se ha logrado una gran cantidad de datos que ha permitido generar modelos que puedan interpretar mejor la información.
- **Estadística y matemáticas:** Es la aplicación de métodos matemáticos para analizar datos.
- **Ciencia de datos:** Es donde se une la estadística y la programación para llevar a cabo los procesamientos de datos.
- **Big data:** Es el conjunto de datos extremadamente grandes y que hace referencia a características como la velocidad, la variabilidad (Datos estructurados y no estructurados) y el volumen.
- **Internet de las cosas:** Es la capacidad de que los dispositivos se comuniquen por medio del internet.
- **Redes neuronales:** Son modelos computacionales inspirados en el funcionamiento del cerebro humano; diseñadas para reconocer patrones y realizar tareas de aprendizaje automático. Al igual que las redes neuronales humanas, tiene las dendritas que son las entradas de información, luego esta es ponderada jerarquizando el conocimiento más importante, se activa y se procesa pasando a otra capa o salida de la información.
- **Compuertas lógicas:** Estas compuertas realizan operaciones lógicas básicas, las cuales son esenciales en el procesamiento de información digital.

Esta se puede basar en la tabla de verdad o tabla de compuerta OR que nos permite bajo algunas condiciones determinar si la salida es 1 (Verdadero) o 0 (Falso) como lo podemos observar en la figura 4 la cual en la única condición donde la salida es falsa es cuando la variable X1 y X2 es 0.

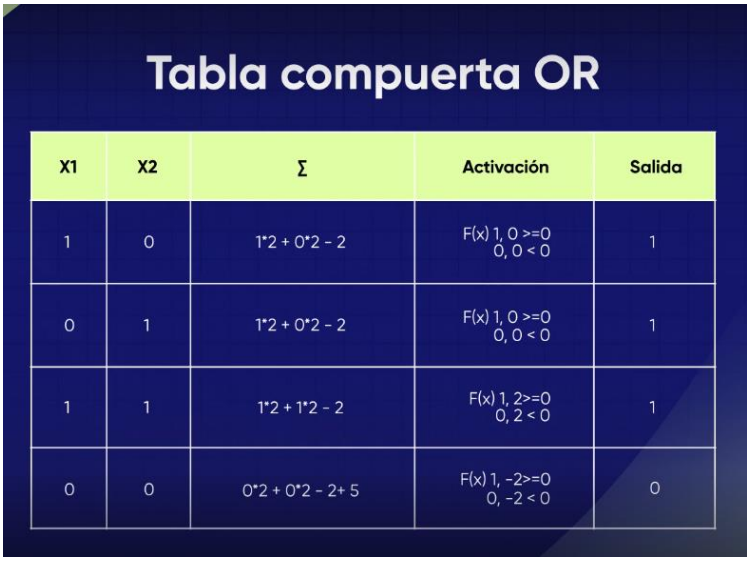

Figura 4: Tabla compuerta OR [47]

Por otro lado, si se evalúa con la tabla de compuerta AND en el único caso donde es verdadero es cuando se cumple tanto para A como para B como varadero y en las demás combinaciones su salida es falso como lo observamos en la figura 5.

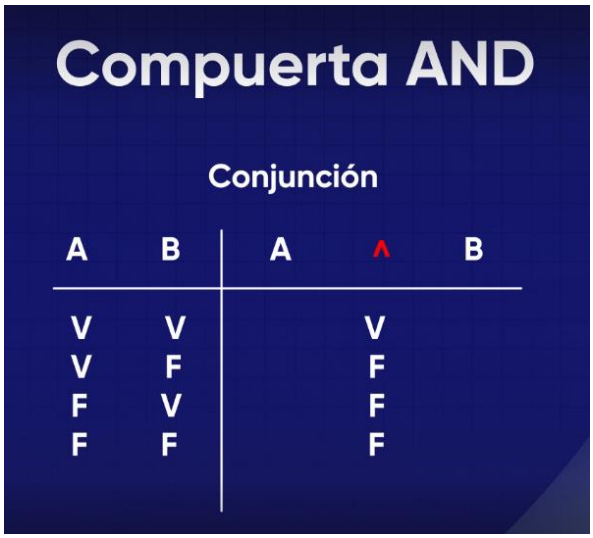

Figura 5: Tabla compuerta AND. [47]

- **Arquitectura de las redes neuronales**: La arquitectura de una red neuronal se refiere a su estructura y organización, incluyendo número de capas, la cantidad de la cantidad de neuronas en cada capa y la conectividad entre ellas. [46]

Dimensionalidad de los datos

La dimensionalidad de los datos se refiere a la cantidad de variables que se tienen en cuenta para cada observación. Para cuando se trabaja con redes neuronales, hay que saber las dimensiones de datos de entrada y salida; cuando se está hablando de un solo dato, se habla de dimensión cero y cuando se habla de dimensión uno, hace referencia a los vectores, también observamos datos en 2D los cuales son los más utilizados y los podemos observar en las tablas de Excel con las filas y las columnas. A medida que los datos son más complejos se van añadiendo dimensiones como por ejemplo imágenes y audios. [48]

Conocimiento jerárquico

Hace referencia a la organización de la información de manera estructurada y jerárquica, donde los conceptos o elementos están dispuestos en niveles ordenados.

Cada capa de la red neuronal tiene su función de activación la cual se compone de función lineal la cual indica que la capa de entrada es igual a la capa de salida. La función umbral o escalonada la cual, si el valor de entrada es menor o igual a 0 la salida será 0, pero si el valor es igual o mayor que 0, la salida será 1. La función logística o sigmoidal interpretada por probabilidad. [49]

Casos de usos de la vida real:

Visión por computadora

Es un campo de la informática que se ocupa de enseñar a las computadoras a interpretar, analizar y entender el contenido visual del mundo; ya sea de imágenes estáticas o de secuencias de video. Estos modelos aún no son muy precisos y aún continúan en desarrollo; es muy importante que se tenga mucha diversidad en los datos para que los modelos generalicen correctamente.

Reconocimiento Biométrico

Reconocimiento facial el cual es un modelo que se va entrenando de acuerdo con los datos que se les va suministrando, por tal motivo es importante la diversidad de datos para evitar sesgos. [50]

Análisis de cluster

Genera agrupaciones de acuerdo con la información que se le suministra al modelo, de esta forma el algoritmo aprende del comportamiento y en este sentido agrupa para luego generar las predicciones. Existen varios algoritmos de clusterización, cada uno con enfoques y características distintas; los más comunes son:

- K-Means: Asigna puntos de datos a k clusters basandoce en la distancia euclidiana entre los puntos y los centroides del cluster.

- Aglomerativo (Hierarchical Clustering): Construye una jerarquía de clusters fusionando gradualmente clusters cercanos. Puede ser representado como un dendrograma.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Agrupa puntos basándose en la densidad de puntos cercanos. Puede identificar clusters de formas arbitrarias y manejar ruido en los datos.
- Mean Shift: Busca modas en la densidad de los datos y asigna puntos hacia estas modas. Puede identificar clusters de formas y tamaños irregulares.
- Clustering Espectral: Utiliza la información del espectro de la matriz de afinidad para realizar la agrupación. Es eficaz para descubrir estructuras complejas y no necesariamente esféricas.
- Gaussian Mixture Model (GMM): Modela los datos como una mezcla de distribuciones gaussianas. Cada punto tiene una probabilidad de pertenecer a cada cluster.
- OPTICS (Ordering Points To Identify Clustering Structure): Similar a DBSCAN, pero proporciona un ordenamiento de la estructura de clusters en lugar de forzar una partición rígida.
- Fuzzy C-Means: Permite que un punto de datos pertenezca a varios clusters con diferentes grados de membresía, en lugar de asignarlos de manera rígida.
- Affinity Propagation:Utiliza mensajes para que los puntos de datos se "autopropongan" como ejemplos representativos y ajusta la asignación basándose en estos mensajes.
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): Construye un árbol de clusters para representar la estructura de clusters de manera eficiente [57]

Evolución del lenguaje natural

Este inicio con los chatbots, los cuales de acuerdo con unas reglas establecidas se generaba un tipo de conversación entre hombre y máquina, sin embargo, no lograba generar una conversación fluida, es allí donde aplicando Machine Learning que por medio de ejemplos le enseña al modelo fue que se logró mejorar el aprendizaje del lenguaje natural en los sistemas y generar interacciones. [51]

GPT-3

Es en la actualidad uno de los algoritmos más potentes, entrenado con más de 165 billones de parámetros y 8 años de recolección de datos. Esta herramienta permite generar predicciones en lenguaje natural muy precisas, sin embargo, puede utilizarse para propósitos pocos éticos como crear noticias falsas por lo que se debe generar reglas en torno a su utilización al igual que cierta cultura en las personas para utilizar responsablemente la herramienta [52]

Github Copilot

Esta herramienta permite trabajar códigos de programación de una manera más sencilla, esta se enlaza con GPT-3 y auto rellena sugiriendo códigos de programación de acuerdo con los que estemos escribiendo similar a cuando nos encontramos escribiendo en un chat. Para lograr esto se ha basado en entrenamiento sobre algoritmos que otros usuarios han escrito. [53]

Introducción a la ética en la inteligencia artificial

La ética en la IA es muy importante debido a varias implicaciones que impactan a la sociedad negativamente y que sin un manejo responsable pueden llegar a generar daños irremediables como en muchas historias pasadas que por irresponsabilidad de muchas personas se ha tenido que enfrentar daño y deterioro en la sociedad por no incluir la ética en los avances de la ciencia y la tecnología. Estas implicaciones incluyen la privacidad de los datos, el sesgo algorítmico, la toma de decisiones automatizada, la transparencia en los modelos y el impacto laboral y social, La ética es necesaria para guiar el desarrollo y utilización de la IA para asegurar que sea justa, transparente y respetuosa con los derechos y valores humanos. En la privacidad de los datos, su recopilación masiva para entrenar modelos de IA plantea muchas preocupaciones sobre los derechos de privacidad individual de los autores. La IA puede desarrollar sesgo que puede resultar en discriminación injusta, en especial en áreas de contratación o decisiones legales. En la toma de decisiones se plantean preguntas éticas sobre quién es responsable en las decisiones erróneas y cómo se puede abordar los prejuicios. La falta de transparencia en algunos modelos de la IA son algunos de los problemas que se pueden enfrentar y es allí donde debe haber un responsable de que estos problemas se corrijan.

La ética en la IA busca equilibrar el progreso tecnológico con la protección de los derechos y valores humanos, promoviendo el desarrollo de tecnologías que beneficien a toda la sociedad, siendo allí donde la educación tiene el potencial de establecer la forma en la tecnología se debe desarrollar. Hay algunas sugerencias que son importantes mencionar para integrar la ética en la IA.

- Considerar la tecnología desde un punto de vista ético y no solo basado en la ciencia.
- Analizar la regulación y documentación que aborden las áreas de riesgo.
- Desarrollar un programa integral de evaluación de riesgos IA que incluya procedimientos, roles, responsabilidades y protocolos.
- Realizar seguimiento a los desarrollos globales y ética de datos.
- Las auditorías basadas en la ética se deben tomar como un proceso continuo y constructivo.

Aplicar la ética en la inteligencia artificial en el contexto turístico y especialmente a las agencias de viajes es esencial para garantizar la confianza, privacidad y el respeto hacia los usuarios, asegurándose de explicar claramente como son utilizados los datos recolectados, evitando el sesgo en las recomendaciones de destinos turísticos, asegurando que las recomendaciones sean equitativas y no estén influenciadas por factores discriminatorios, asegurar que los sistemas sean seguros y resistentes ante ataques

cibernéticos; enseñando a los usuarios como llega una recomendación para construir lazos de confianza entre las agencias y los clientes, asegurando que los servicios prestados por la agencia sean accesibles para todas las personas, evaluando el impacto social de los sistemas IA considerando como afecta a las comunidades locales y a la sostenibilidad del turismo en diferentes destinos, realizando pruebas para detectar y mitigar posibles sesgos en los algoritmos. [58]

Es importante considerar la ética en las prácticas de inteligencia artificial para contribuir a un turismo más responsable y respetuoso con la sociedad.

Caso de éxito en la industria de la IA

Local Interpretable Model-agnostic Explanations: Este modelo fue propuesto por Marco Tulio Riveiro en el 2016 para explicar modelos de Machine Learning a través de tomar la versión original y aplicar el modelo complejo a estas instancias perturbadas para luego ajustándolo a un modelo interpretable a los resultados obtenidos, proporcionando una explicación más comprensible de la predicción del modelo complejo.

Como conclusión general es importante destacar que la ética depende de todas las personas, ya sea quienes desarrollan las tecnologías y de quienes la utilizan ya que la IA por sí sola no tiene la capacidad inherente de darse cuenta de sus prejuicios y crecer a partir de ellos [54]

Innovación tecnológica con inteligencia artificial.

La inteligencia artificial de por sí ya es una innovación tecnológica dado a que trata de recrear el comportamiento de nuestro cerebro para diversos propósitos, sin embargo, aún falta por perfeccionarse. Aún falta desarrollarse en aspectos biológicos para entender el comportamiento humano y de allí brindar una respuesta adecuada a los diferentes estímulos, también en lo filosófico con respecto a la ética que se mencionó anteriormente.

Para que los modelos de Inteligencia Artificial tengan éxito es necesario entender el problema que se quiere abordar y atacar desde diferentes puntos de vista con el fin de brindar la mejor solución ante diferentes situaciones. [55]

En ocasiones pueden surgir problemas de investigación que requieren un tratamiento más rápido para la obtención de información que es de interés para abordar el problema que aplicar métodos tradicionales para obtener dicha información. Para ello se puede utilizar métodos en la red que permite la obtención de esta información, es necesario como se ha mencionado anteriormente la importancia de entender el negocio para determinar qué información da valor.

Lo que se busca en la agencia de turismo aplicando la innovación tecnológica con inteligencia artificial es la de personalizar a nuestros clientes, personalizando las atenciones para cada uno de ellos de acuerdo con su comportamiento de compra y las

preferencias que cada uno tiene. Esto nos brindara la oportunidad de generar las recomendaciones adecuadas como destinos, ofertas y actividades.

Para los clientes es importante los costos, de acuerdo con el comportamiento del mercado podremos determinar la demanda y generar las mejores ofertas para nuestros clientes.

Como se mencionó anteriormente, la obtención de la información por medio de la red, en este caso el de las redes sociales es vital para conocer los comentarios de los clientes, para ellos es necesario la implementación del lenguaje natural que nos permita de una forma más sencilla extraer la información y tomar decisiones.

Por último, debemos contar con sistemas de seguridad en nuestras bases de datos, a fin de proteger la base de datos de los clientes y evitar amenazas que puedan generar estafas a nuestros clientes.

Algoritmo evolutivo

Ayuda a optimizar herramientas a partir de simulaciones, este no tiene datos sino entornos que permiten evolucionar y encontrar soluciones óptimas.

Lógica difusa

Es un modelo que ayuda a desarrollar modelos con parámetros lingüísticos, este se utiliza en casos donde no se requiere una respuesta numérica si no con etiquetas lingüísticas.

Como se ha mencionado anteriormente el aprendizaje automático se divide en 3, un aprendizaje supervisado que nos brinda una respuesta binaria, un aprendizaje no supervisado que los datos no tienen un orden como en el supervisado y el aprendizaje por refuerzo el cual se basa en un entorno donde el modelo realiza diferentes pruebas y se utiliza un sistema de recompensas para que cada que se haga bien el pronóstico nos acercamos a la solución, en caso de que se haga mal, se castiga y ayuda a re entrenar el algoritmo. [56]

Desarrollo e implementación del aprendizaje

En un mundo cada vez más interconectado, las empresas buscan comprender y anticipar las necesidades de sus clientes para proporcionar experiencias personalizadas y satisfactorias. En este contexto, el siguiente proyecto se centra en el uso de técnicas de Machine Learning para abordar la pregunta fundamental de investigación: ¿Cuál es la probabilidad de que una persona que compre un paquete turístico por primera vez con la agencia de viajes Agentur, desee realizar otras futuras compras en la misma agencia aplicando algoritmos de predicción-regresiones?

Para abordar este proyecto, nos enfocamos en un caso de estudio concreto: la agencia de viajes denominada "Agentur". El objetivo principal es anticipar la disposición de un cliente a realizar compras adicionales con la misma agencia. A través del análisis de una tabla de datos detallada, examinamos las probabilidades asociadas con la repetición de compras de paquetes turísticos. Exploramos no solo la posibilidad de que un cliente adquiera un segundo paquete, sino también su propensión a repetir la experiencia para paquetes adicionales. Este enfoque integral incluye la evaluación de patrones de interés, niveles de satisfacción, preferencias por destinos específicos, y la elección de categorías hoteleras. A través de este análisis detallado, buscamos desentrañar las complejidades que influyen en la decisión de los clientes, proporcionando una perspectiva valiosa para mejorar las estrategias de la agencia "Agentur" y fortalecer su relación con los clientes.

Como se menciona en los conceptos introductorios, es muy importante seguir la secuencia de la pirámide de valor desde la base para lograr el objetivo del proyecto. En este sentido, la agencia de viajes debe comenzar por realizar una inversión en tecnología; por lo tanto, se solicita, antes de dar inicio al proyecto de aprendizaje automático, contar con los siguientes equipos tecnológicos y recursos:

1. Hardware Potente:
   Una computadora con recursos significativos, preferiblemente con una unidad de procesamiento grafico (GPU) potente, ya que muchas tareas de machine learning se benefician enormemente de la aceleración GPU.
2. Entorno de desarrollo:
   Herramientas como Python y Jupyter Notebook son comúnmente utilizadas en el desarrollo de proyectos de machine learning. Tambien es útil contar con bibliotecas populares como Tensorfloww o PyTorch para implementar modelos de manera efectiva.
3. Conjuntos de Datos Relevantes: Se necesitarán conjuntos de datos extraídos de la herramienta XMART para entrenar y evaluar el modelo de machine learning.
4. Software de Visualización de datos: Herramientas como Matplotlib o Seaborn en Python son útiles para visualizar datos y resultados.
5. Plataformas en la nube: Servicios en la nube como AWS, Google cloud Platform (GCP) o Microsoft Azure proporcionan recursos escalables para entrenar modelos en grandes conjuntos de datos.
6. Control de versiones: Utilizar sistemas de control de versiones con Git para llevar un registro de los cambios en el código y colaborar eficientemente en equipo.

7. Documentación: Plataformas de documentación como Jopyter Notebooks, junto con buenas prácticas de documentación en código, son esenciales para comprender y compartir el trabajo.
8. Bases de datos: En algunos casos, puede ser necesario un sistema de gestión de bases de datos para almacenar y recuperar datos de manera eficiente.
9. Conjunto de herramientas de desarrollo: Puede ser útil contar con la herramienta de desarrollo IDEs y editores de código.

Ya cuando nos aseguramos de contar con las herramientas tecnológicas necesarias, continuamos con la implementación gobernanza de datos, la cual es fundamental para garantizar la calidad e integridad de los mismos; esto implica segur un conjunto de pasos específicos adaptados a las necesidades de la agencia de viajes “Agentur” que se describen a continuación:

1. Evaluación de la Situación Actual:
- Realiza una evaluación exhaustiva de la situación actual de la gestión de datos en "Agentur". Identificar los activos de datos críticos, los procesos existentes y las áreas que necesitan mejoras.

2. Establecer un Equipo de Gobernanza de Datos:
- Designe un Chief Data Officer (CDO) o un líder de gobernanza de datos responsable de liderar la implementación.
- Forma un equipo de gobernanza de datos que incluye representantes de diferentes departamentos para garantizar una perspectiva holística.

3. Definir Políticas y Normativas:
- Desarrolla políticas y normativas claras que abordan la recopilación, almacenamiento, acceso y uso de datos en "Agentur". Asegúrese de cumplir con las regulaciones de privacidad y protección de datos.

4. Inventario y Catalogación de Datos:
- Crea un inventario detallado de todos los activos de datos relevantes para la agencia de viajes, incluyendo datos de clientes, itinerarios, transacciones y preferencias.
- Establece un catálogo de metadatos que describe la naturaleza y el uso de cada conjunto de datos.

5. Asegurar Calidad de Datos:
- Implementa estándares de calidad para garantizar la integridad y precisión de los datos.
- Establece procesos regulares de limpieza y validación de datos

6. Seguridad de Datos:
- Definir niveles de acceso y control para proteger datos sensibles.

- Implemente medidas de seguridad robustas, como cifrado y monitoreo continuo de actividades.

7. Gestión de Cambios y Actualizaciones:
- Establece un proceso formal para la gestión de cambios en la estructura y definición de datos.
- Implementa un sistema de aprobación para cambios significativos.

8. Capacitación y Concienciación:
- Ofrece programas de formación sobre las políticas y prácticas de gobernanza de datos para todos los empleados.
- Fomenta una cultura de concienciación sobre la importancia de la gestión adecuada de datos.

9. Colaboración Interdepartamental:
- Establece comités de gobernanza de datos que incluyen representantes de diferentes áreas para facilitar la colaboración.
- Colabora estrechamente con departamentos como marketing, ventas y atención al cliente para asegurar la coherencia de los datos.

10. Documental y Comunicar:
- Documenta todas las políticas y procedimientos de gobernanza de datos.
- Comunica periódicamente las actualizaciones y cambios en las políticas de gobernanza de datos a todo el personal.

Al implementar estos pasos, "Agentur" puede mejorar significativamente la calidad de sus datos, fortalecer la seguridad de la información y optimizar el uso de los datos para tomar decisiones informadas en la gestión de sus servicios de viajes. [59]

Para predecir la probabilidad de que una persona que compra por primera vez en la agencia de viajes “Agentur” vuelva a utilizar la agencia para realizar futuras compras, se considera utilizar la herramienta Machine Learning, específicamente el modelo de regresión logística, se utiliza este algoritmo y no el de regresión lineal dado a que este brinda un resultado continuo y cuantitativo y lo que se busca es que nos arroje un resultado binario como lo hace el algoritmo de regresión logística, en este caso que prediga la probabilidad de que un cliente realice compras futuras, es decir si o no repite compra. Utilizando la tabla de datos disponible, este modelo evaluará diversas variables, como patrones de interés, niveles de satisfacción, repeticiones de destinos, preferencias hoteleras, entre otros, para estimar la probabilidad de retención del cliente.
Al entrenar y validar este modelo con conjuntos de datos históricos, obtendremos una herramienta predictiva capaz de clasificar a los clientes en categorías de probabilidad. Con esta información, la agencia podrá segmentar su base de clientes, personalizar estrategias de retención y marketing, diseñando ofertas específicas para maximizar la fidelización. La interpretación de los coeficientes del modelo también proporciona una

comprensión más profunda sobre qué factores tienen un impacto significativo en la probabilidad de repetición de compras.
En resumen, la respuesta a este problema implica la aplicación de técnicas de Machine Learning para desarrollar un modelo predictivo que permita a la agencia “Agentur” anticipar y abordar de manera proactiva las necesidades de sus clientes, mejorando así su capacidad para retener y satisfacer las necesidades de forma efectiva.

Para la aplicación del modelo de regresión logística de Machine Learning se recomienda la utilización de Python para el desarrollo del algoritmo. Se busca con este modelo el de predecir si un cliente repite compra o no y la tomaremos como nuestra variable dependiente. (1 si el cliente repite compra, 0 si no lo hace).
Al aplicar Python se requieren de algunas librerías que nos permitirá realizar el entrenamiento, test y evaluar el modelo con las métricas de desempeño mencionadas durante el contexto del informe.

Para aplicar el modelo se utilizó la siguiente base de datos, en la tabla 1 podemos observar parte de ella con las variables más significativas y que se utilizaron para el desarrollo del modelo. Esta cuenta con 12 variables incluyendo la variable objetivo y 2101 observaciones.

Tabla 1. Base de datos para la aplicación del modelo de Machine Learning

NOMBRE	CATEGORIA HOTEL	ASESOR	LINEA AEREA PREFERENTE	OFICINA_ASESOR	ESTAD	FECHA_CREACIO	CIUDAD	REPITE COMPRA
ABIGAIL ESCOBAR DE ARENAS	4 ESTRELLAS	Mercedes Garcia Baez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	18/8/2023	Bogota	NO
ADAN GIRALDO PULGARIN	5 ESTRELLAS	Felipe Andres Ocampo Jaramillo	AVIANCA	VIAJES AGENTUR SA - OFICINA PEREIRA	Activo	14/3/2022	Pereira	NO
ADRIANA ARISTIZABAL VILLA	4 ESTRELLAS	Margarita Maria Giraldo Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	27/4/2022	Manizales	NO
ALEY MONTILVA MUNOZ	5 ESTRELLAS	Marcela Arango Villegas	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	10/10/2022	Bogota	SI
ADRIANA DIAZ BETANCOURT	5 ESTRELLAS	Margarita Maria Giraldo Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	13/6/2022	Barranquilla	NO
ADRIANA GOMEZ BERNAL	4 ESTRELLAS	Dora Isabel Orozco Vallejo	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	13/1/2023	Cali	NO
ADRIANA GONZALEZ PINEROS	5 ESTRELLAS	Sofia Gomez Londono	AVIANCA	VIAJES AGENTUR SA - OFICINA BOGOTA	Activo	19/5/2023	Cucuta	NO
ADRIANA ISABEL HERRERA MARIN	4 ESTRELLAS	Mercedes Garcia Baez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	12/8/2022	Bucaramanga	NO
ESPERANZA HURTADO OCAMPO	5 ESTRELLAS	Dora Isabel Orozco Vallejo	LATAM	VIAJES AGENTUR SA - OFICINA CABLE	Activo	24/1/2022	Manizales	SI
ADRIANA ISAZA NARVAEZ	5 ESTRELLAS	Mercedes Garcia Baez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	2/11/2023	Manizales	NO
ADRIANA JARAMILLO GIRALDO	4 ESTRELLAS	Margarita Maria Giraldo Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	12/1/2022	Manizales	NO
ADRIANA JARAMILLO MEJIA	5 ESTRELLAS	Sofia Gomez Londono	AVIANCA	VIAJES AGENTUR SA - OFICINA BOGOTA	Activo	24/4/2023	Pereira	NO
ADRIANA MABEL LOPEZ RODRIGUEZ	4 ESTRELLAS	Felipe TOURLINK	AVIANCA	VIAJES AGENTUR SA - TOURLINK	Activo	31/12/2022	Bogota	NO
ADRIANA MARCELA GORDILLO GONZALEZ	5 ESTRELLAS	Margarita Maria Giraldo Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	5/4/2023	Manizales	NO
ADRIANA MARGARITA TREJOS ATEHORTUA	4 ESTRELLAS	Margarita Maria Giraldo Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	23/8/2023	Cali	NO
ADRIANA MARIA BUENO MUÑOZ	5 ESTRELLAS	Felipe TOURLINK	AVIANCA	VIAJES AGENTUR SA - TOURLINK	Activo	13/10/2022	Manizales	NO
ADRIANA MARIA CANO SALAZAR	4 ESTRELLAS	Lina Maria Buitrago Lopez	AVIANCA	VIAJES AGENTUR SA - OFICINA BOGOTA	Activo	26/1/2022	Pereira	NO
ADRIANA MARIA DUQUE ARISTIZABAL	5 ESTRELLAS	Nathaly Moreno Gomez	AVIANCA	VIAJES AGENTUR SA - OFICINA CABLE	Activo	24/1/2022	Bogota	NO

Códigos de aplicación del modelo

A continuación, se describen los códigos de programación a utilizar para este proyecto:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, roc_curve, auc

# Cargar datos
data = pd.read_csv('CLIENTES.csv')

# La columna llamada 'REPITE_COMPRA' indica si el cliente repite la compra (1) o no (0), esta será la variable objetivo
# El resto de las columnas son características a utilizar para predecir la repetición de la compra.

# Dividir en características (X) y variable dependiente (y)
X = data[['NOMBRE', CATEGORIA_HOTEL, 'ASESOR', 'LINEA_AEREA_PREFERENTE', 'OFICINA_ASESOR', 'FECHA_CREACION','CIUDAD']]
y = data['REPITE_COMPRA']

# Dividir en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Escalar características
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Crear modelo de regresión logística
model = LogisticRegression()

# Entrenar el modelo
model.fit(X_train_scaled, y_train)

# Realizar predicciones en el conjunto de prueba
y_pred = model.predict(X_test_scaled)

# Evaluar el rendimiento del modelo
print(f'Accuracy: {accuracy_score(y_test, y_pred)}')
print('\nConfusion Matrix:')
print(confusion_matrix(y_test, y_pred))
print('\nClassification Report:')
print(classification_report(y_test, y_pred))

# Obtener la matriz de confusión
cm = confusion_matrix(y_train, y_train_pred)

# Configurar el estilo del mapa de calor
sns.set(font_scale=1.2)
```

```
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Repite', 'Repite'],
yticklabels=['No Repite', 'Repite'])

# Configurar etiquetas y título
plt.title('Matriz de Confusión - Datos de Entrenamiento')
plt.xlabel('Predicciones')
plt.ylabel('Valores Verdaderos')

# Mostrar el mapa de calor
plt.show()

# Crear un gráfico de dispersión de las características con la probabilidad predicha
plt.figure(figsize=(12, 6))

# Seleccionar solo dos características para visualizar (puedes ajustar según tus
necesidades)
feature1 = ' NOMBRE '
feature2 = 'FECHA_CREACION'

# Scatter plot de los datos de entrenamiento
plt.scatter(X_train_scaled[y_train == 0][feature1], X_train_scaled[y_train ==
0][feature2], label='No Repite Compra', alpha=0.5)
plt.scatter(X_train_scaled[y_train == 1][feature1], X_train_scaled[y_train ==
1][feature2], label='Repite Compra', alpha=0.5)

# Líneas de decisión del modelo
x_min, x_max = X_train_scaled[:, 0].min() - 1, X_train_scaled[:, 0].max() + 1
y_min, y_max = X_train_scaled[:, 1].min() - 1, X_train_scaled[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.1), np.arange(y_min, y_max, 0.1))
Z = model.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
plt.contourf(xx, yy, Z, alpha=0.3, cmap='coolwarm')

plt.xlabel(feature1)
plt.ylabel(feature2)
plt.title('Comportamiento de Datos con Probabilidad Predicha (Entrenamiento)')
plt.legend()
plt.show()

# Crear la curva ROC para evaluar el rendimiento del modelo
y_pred_prob = model.predict_proba(X_test_scaled)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)

plt.figure(figsize=(8, 8))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='Curva ROC (AUC = %0.2f)' % roc_auc)
```

```
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('Tasa de Falsos Positivos')
plt.ylabel('Tasa de Verdaderos Positivos')
plt.title('Curva ROC')
plt.legend(loc="lower right")
plt.show()
```

Resultados del modelo

Una vez aplicado el modelo de regresión logística lo que se busca es predecir el comportamiento que tiene los clientes en la decisión de repetir la compra en la agencia de viajes, de esta forma poder determinar estrategias de publicidad más personalizadas tanto para los clientes que repitieron la compra como los que no lo hicieron. Teniendo en cuenta los resultados es importante enfocarnos en aquellos clientes que no repitieron compra dado a que se pueden tomar estrategias de retención con el fin de que ejecuten una segunda compra.

Para que el modelo tenga éxito es importante tener en cuenta la experiencia de aquellos asesores que llevan más tiempo con la agencia y determinar nuevas variables que puedan ser determinantes para predecir la variable objetivo, dado a que como este modelo apenas se va a implementar los datos que se tienen hasta el momento pueden no ser relevantes pero da como partida la construcción de nuevas variables de interés que permita tomar mejores estrategias publicitarias de acuerdo con las predicciones que el modelo nos arroja.

La aplicación del algoritmo no debe quedarse solo con los datos que le indicamos en su primera aplicación, sino también de ir mejorándolo con las nuevas variables y actualización constante de la base de datos. Para ellos es importante generar dentro de los algoritmos del modelo las gráficas de tendencias que nos muestre el comportamiento de los datos, esto se puede lograr gracias a las librerías de matplotlib donde podemos generar histogramas, diagramas de barras. También con la aplicación de la matriz de correlación y en la evaluación del modelo con la curva de ROC. Hay diferentes algoritmos mencionados anteriormente que nos permite imprimir las gráficas y visualizar el comportamiento de los datos de una forma más intuitiva.

La implementación de Machine Learning y principalmente de la inteligencia artificial, permitirá ser más competitivos en el mercado donde plataformas como Booking ya ejecuta predicciones de acuerdo al comportamiento de navegación de las personas y de esta manera hace recomendaciones de viajes. Se busca llegar a este nivel lo cual traerá un crecimiento económico en la agencia gracias a la correcta utilización de la tecnología y apoyado en un recurso importante como lo son los datos.

Conclusiones

En este proyecto fue posible aprender conceptos en torno a la inteligencia artificial, a su historia y cómo funciona, la importancia del Machine Learning para predecir comportamientos y generar recomendaciones de acuerdo con el modelo que estemos construyendo. Vimos como grandes empresas como Netflix, Mercado Libre, Amazon y Spotify tienen éxito en la aplicación de los modelos y que permiten generar valor a los datos para facilitar a los usuarios su navegación en las plataformas.

Como estudiantes de ingeniería industrial, fue de gran valor tener un primer acercamiento con las nuevas tecnologías, gracias a esto se generaron intereses de aprender el lenguaje de programación como base para la aplicación de modelos de Machine Learning. Aunque para nosotros algunos conceptos no fueron claros dado al nivel que se requería, fue posible entender su base y las principales herramientas que se utilizan, como las librerías y herramientas para el desarrollo del código. Así mismo la importancia que tiene los datos para poder generar los modelos. También es necesario conocer el problema y hacerse las preguntas correctas que permitan encontrar la mejor solución.

Para el desarrollo del proyecto y una vez conociendo los conceptos más relevantes fue posible aplicar un modelo de regresión lineal que nos permite predecir la probabilidad de repetición de compra de un paquete turístico en una agencia de viajes, para ello se utilizaron variables como nombre, categoría del hotel, asesor que hizo la venta, línea aérea preferente por el cliente, oficina de venta, ciudad y la variable objetivo que en este caso era si repetía o no una compra.

Aunque el modelo no fue desarrollado propiamente en un ambiente de Python, las métricas de desempeño elegidas tanto para los datos de prueba como de entrenamiento se estima en valores superiores al 85% y que nos garantiza un funcionamiento óptimo del modelo ante el ingreso de nuevos datos. La matriz de confusión bajo su esquema nos permitirá observar la proporción falsos positivos y falsos negativos la cual se espera que sea baja.

Con la aplicación del modelo permitirá generar estrategias de negocio que pueda retener a los clientes existentes y atraer a nuevos clientes, analizando su comportamiento de compra podremos predecir si el cliente realizará una nueva compra y en base a esto ofrecerle nuevos planes turísticos que sean atractivos. La ejecución de atenciones más personalizadas generará mayor valor a la agencia lo que se traduce en mayores ingresos y más competitividad en el mercado.

Referencias

[1] J. Martínez, "Inteligencia Artificial", en computerworld.es, 2017.

[2] L.A. Lee, "Machine Learning en tu día a día", en Introducción a Machine Learning, 2021.
[3] L.A. Lee, "La data a veces es sexy", en Introducción a Machine Learning, 2021.
[4] L.A. Lee, "Desmitificando", en Introducción a Machine Learning, 2021.
[5] L.A. Lee, "Estrategia de Datos", en Introducción a Machine Learning, 2021.
[6] L.A. Lee, "Retando a los expertos", en Introducción a Machine Learning, 2021.
[7] L.A. Lee, "Clasificación vs. Regresión", en Introducción a Machine Learning, 2021.
[8] L.A. Lee, "La básica y confiable, la regresión lineal", en Introducción a Machine Learning, 2021.
[9] L.A. Lee, "Los tipos de datos: ¿Tienen estructura?", en Introducción a Machine Learning, 2021.
[10] L.A. Lee, "Python, Paquetes y librerías", en Fundamentos aplicados de Machine Learning, 2021.
[11] J.A. Rodrigo, "Ejemplo matriz de correlaciones", en Correlación lineal con Python, 2020. URL: https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html
[12] L.A. Lee, "Prediciendo el precio de una casa: Regresiones en Boston", en Fundamentos aplicados de Machine Learning, 2021.
[13] L.A. Lee, " ¿A quién le prestarías dinero? Clasificación para tarjetas de crédito", en Fundamentos aplicados de Machine Learning, 2021.
[14] L.A. Lee, "El Ciclo de Vida del Proyecto", en Fundamentos aplicados de Machine Learning, 2021.
[15] H.A. Aragón, "Inteligencia artificial, Machine Learning supervisado y no supervisado", en Machine Learning: Aprendizaje supervisado, 2021.
[16] H.A. Aragón, "Métricas de desempeño", en Machine Learning: Aprendizaje supervisado, 2021.
[17] H.A. Aragón, "Ejemplos de métricas de desempeño", en Machine Learning: Aprendizaje supervisado, 2021.
[18] H.A. Aragón, "Data set, train, test and evaluation", en Machine Learning: Aprendizaje supervisado, 2021.
[19] H.A. Aragón, "Regresión lineal", en Machine Learning: Aprendizaje supervisado, 2021.
[20] H.A. Aragón, "Regresión logística", en Machine Learning: Aprendizaje supervisado, 2021.
[21] H.A. Aragón, "Árboles de decisión - continuos", en Machine Learning: Aprendizaje supervisado, 2021.
[22] H.A. Aragón, "Random forest", en Machine Learning: Aprendizaje supervisado, 2021.
[23] H.A. Aragón, "Introducción neural network", en Machine Learning: Aprendizaje supervisado, 2021.
[24] C. Tabares, "¿Qué es un contrafactual?", en Machine Learning: Análisis contrafactual, 2022.
[25] C. Tabares, "Correlacción vs. causalidad", en Machine Learning: Análisis contrafactual, 2022.
[26] C. Tabares, "Variables omitidas y sesgo de selección", en Machine Learning: Análisis contrafactual, 2022.

[27] C. Tabares, "Experimentos estratificados", en Machine Learning: Análisis contrafactual, 2022.
[28] C. Tabares, "Repaso de aprendizaje de máquina", en Machine Learning: Análisis contrafactual, 2022.
[29] C. Tabares, "Inferencia causal con Big Data", en Machine Learning: Análisis contrafactual, 2022.
[30] C. Tabares, "Double LASSO", en Machine Learning: Análisis contrafactual, 2022.
[31] C. Tabares, "Causal Trees", en Machine Learning: Análisis contrafactual, 2022.
[32] C. Tabares, "Causal Forest", en Machine Learning: Análisis contrafactual, 2022.
[33] M. Rojo, "Machine Learning Workflow", en Evaluación de modelos de Machine Learning, 2023.
[34] M. Rojo, "Preparación de datos", en Evaluación de modelos de Machine Learning, 2023.
[35] M. Rojo, "Matriz de confusión", en Evaluación de modelos de Machine Learning, 2023.
[36] M. Rojo, "Modelado", en Evaluación de modelos de Machine Learning, 2023.
[37] L. Gonzales, "Matriz de confusión", en Curso Machine Learning con Python, 2019. URL: https://www.youtube.com/watch?v=r5WIImKV1XA&t=2s
[38] M. Rojo, "Accuracy", en Evaluación de modelos de Machine Learning, 2023.
[39] M. Rojo, "Precision y Recall", en Evaluación de modelos de Machine Learning, 2023.
[40] M. Rojo, "Specificity y F1 score", en Evaluación de modelos de Machine Learning, 2023.
[41] M. Rojo, "Curva ROC y AUC", en Evaluación de modelos de Machine Learning, 2023.
[42] Hospital Universitario Ramón y Cajal, "Tipos de curvas de roc", en Curvas ROC, S.A. URL: http://www.hrc.es/bioest/roc_1.html
[43] M. Rojo, "Curva de Precision-Recall", en Evaluación de modelos de Machine Learning, 2023.
[44] M. Rojo, "Métodos de validación cruzada", en Evaluación de modelos de Machine Learning, 2023.
[45] F. Ruiz, "Principios del machine Learning", en Introducción a la Inteligencia Artificial, 2022.
[46] F. Ruiz, "Conceptos alrededor de la inteligencia artificial", en Introducción a la Inteligencia Artificial, 2022.
[47] F. Ruiz, "Compuertas lógicas", en Introducción a la Inteligencia Artificial, 2022.
[48] F. Ruiz, "Dimensionalidad de los datos", en Introducción a la Inteligencia Artificial, 2022.
[49] F. Ruiz, "Conocimiento jerárquico", en Introducción a la Inteligencia Artificial, 2022.
[50] F. Ruiz, "Visión por computadora", en Introducción a la Inteligencia Artificial, 2022.
[51] F. Ruiz, "Evolución del NLP", en Introducción a la Inteligencia Artificial, 2022.
[52] F. Ruiz, "GPT-3", en Introducción a la Inteligencia Artificial, 2022.
[53] F. Ruiz, "Github Copilot", en Introducción a la Inteligencia Artificial, 2022.

[54] E. Wohlmuth, "Estrategias responsables y prácticas transparentes en el uso de la IA", en Introducción a la ética en la Inteligencia Artificial, 2023.
[55] R.A. Murga, "Beneficios de la IA en la innovación tecnológica", en Innovación tecnológica con inteligencia artificial, 2023.
[56] R.A. Murga, "Diseño y desarrollo de soluciones con IA - Parte II", en Innovación tecnológica con inteligencia artificial, 2023.
[57] O. Fernandez, "El análisis de clusters, aplicación, interpretación y validación", en file:///C:/Users/red/Downloads/25102-Text%20de%20l'article-58670-1-10-20061130.pdf.2024
[58] E. Wohlmuth, "Estrategias responsables y prácticas transparentes en el uso de la IA", en Introducción a la ética en la Inteligencia Artificial, 2023.
[59] E. Rojas, "Machine Learning", en análisis de lenguajes de programación y herramientas para desarrollo. 2024.