El Machine Learning En Academias de Educación Vial

Corporación Universitaria Remington.
Facultad de Ingeniería
Tecnología en Desarrollo de Software.

Estudiantes:
Niyileth Karina Enriquez Quintero.
Laura Valentina Bautista.
Instructor:
Juan Pablo Vélez Uribe.
Seminario.
2023.

Dedicatoria

A nuestras familias.

A nuestros mentores.

A todos aquellos que creyeron en nosotras.

Agradecimientos

A nuestras familias, por su apoyo incondicional y compresión durante todo el proceso, han sido ese impulso que muchas veces necesitamos para poder seguir adelante, a pesar de las adversidades, han sido nuestro soporte en el cual poco a poco le vamos reflejando nuestros cambios y mejoras.

A nuestros mentores por contribuir como guías en todo este proceso, compartiéndonos sus conocimientos y dándonos las herramientas necesarias para hacer de este trabajo algo eficiente, les estaremos siempre agradecidas.

A todos los que creyeron en nosotras, ya que nos brindaron motivación y colaboración en cada paso de este proceso, sin su ayuda, esto no hubiera sido posible, gracias a todo corazón por formar parte de este logro

Tabla de Contenido

RESUMEN

En este proyecto, se exploraron varios temas relacionados con la evaluación y supervisión de procesos en sistemas de información apoyados por el Aprendizaje Automático. Se aplicó el Aprendizaje Automático en Academias de Educación Vial para mejorar la gestión de la enseñanza y la evaluación de los conductores, centrándose en la predicción del consumo de combustible de los vehículos de enseñanza para optimizar los recursos y reducir los costes.

La introducción al Aprendizaje Automático destacó su capacidad de aprender sin programación explícita, utilizando datos para mejorar continuamente el rendimiento del sistema. Se abordaron conceptos como el aprendizaje supervisado y no supervisado, junto con la importancia de la Ciencia de Datos. En el análisis de datos de las Academias de Educación Vial, se exploró cómo el Aprendizaje Automático puede identificar patrones en el rendimiento de los alumnos, personalizando la instrucción y contribuyendo a la automatización, especialmente en el control eficiente del combustible.

Se presentó el concepto de Big Data como principal insumo para las aplicaciones de Aprendizaje Automático, generando información valiosa para la toma de decisiones estratégicas. También se hizo hincapié en la importancia de una buena estrategia de datos y en la pirámide de valor de los datos.

En el mundo de la ciencia de datos, el Aprendizaje Automático se ha convertido en un cambio de juego, revolucionando la forma en que abordamos problemas complejos. Sin embargo, con la creciente complejidad de los modelos, la necesidad de interpretabilidad e inferencia causal se ha vuelto más crítica que nunca.

Para hacer frente a esto, se han desarrollado varios algoritmos, como Propensity Score, Double LASSO, Causal Trees y Causal Forest, para mejorar la interpretación de los modelos y permitir una toma de decisiones informada. Estos algoritmos se centran en comprender la causalidad y la correlación, que son esenciales para desarrollar modelos sólidos.

El proceso de evaluación de los modelos de Aprendizaje Automático es un paso crucial en este ámbito. Implica comprender el problema empresarial, preparar y modelar los datos, evaluar el rendimiento del modelo y finalizarlo. La preparación de los datos

implica utilizar herramientas como pandas en Python, emplear técnicas como MinMaxScaler y get_dummies, y asegurarse de que los datos están limpios y listos para el análisis.

Se hace hincapié en el entrenamiento de modelos y la selección de algoritmos, junto con métricas de rendimiento como las matrices de confusión, la exactitud, la precisión, el recuerdo y otras. Las técnicas de validación cruzada son esenciales para una sólida selección de modelos, y el aprendizaje supervisado mediante conjuntos de datos etiquetados se utiliza para entrenar algoritmos.

Palabras clave

Machine Learning, Inteligencia artificial, Aprendizaje Supervisado, Aprendizaje no Supervisado, Python, Rstudio, Innovación tecnológica, Big Data, Pirámide de Valor de Los Datos, Variable Objetivo, Regresión, Clasificación, Academia de Enseñanza Automovilística, Vehículo de Enseñanza.

I. MARCO CONCEPTUAL Y CONTEXTUAL

Dentro la elaboración de este proyecto, se vieron muchas temáticas que evaluaban y supervisaban varios procesos dentro de los sistemas de información, que eran respaldados y ejecutados de la mano del Machine Learning, con el cual hoy en día se ha hecho uso en muchos ámbitos de la vida cotidiana, en nuestro caso, el Machine Learning lo estamos aplicando específicamente a las **Academias de Educación Vial**, debido a que con este podemos denotar su uso ampliamente en áreas como la inteligencia artificial, asistentes virtuales y hasta en análisis de datos para la generación de decisiones automatizadas. Nuestro motivo principal con el uso y apoyo del **Machine Learning,** es buscar que academias de conducción que no se han actualizado en el uso de procesos modernos para la gestión de sus compañías, opten por esta mejora, y, sobre todo, puedan tener un control eficiente de los vehículos de enseñanza, que en muchas ocasiones no tienden a regularles el consumo de combustible, y en esos momentos se ve gastos más inoportunos e innecesarios.

Para abordar el problema específico en la empresa de enseñanza de conducción, se consideró el consumo de combustible de cada vehículo usado por los instructores según varias variables, como tipo de vehículo, distancia recorrida, condiciones de conducción y comportamiento histórico. La variable de interés sería la cantidad de combustible que se consumirá en una determinada ruta o período de tiempo. Esta predicción implicaría un ejercicio de regresión, ya que estamos buscando prever un valor numérico (cantidad de combustible) en lugar de clasificar en categorías. La información requerida para este propósito constituiría datos estructurados, ya que tendríamos registros históricos organizados en tablas con variables cuantitativas y cualitativas. El objetivo de predecir el consumo de combustible permitiría tomar decisiones informadas sobre la asignación de vehículos y rutas, optimizando así los recursos y reduciendo los costos operativos asociados al combustible.

A. INTRODUCCIÓN A MACHINE LEARNING

Cuando se aborda la temática del Machine Learning, se enfoca en un hito significativo en la evolución tecnológica: capacitar a máquinas para aprender sin programación explícita. En su esencia, este enfoque busca dotar a los sistemas de la capacidad de aprender a partir de datos, contribuyendo a mejoras continuas en su rendimiento a lo largo del tiempo. Esto nos proporciona un enfoque revolucionario para abordar problemas complejos y tomar decisiones.

Este comienza su proceso y ejecución con la alimentación de sus algoritmos con datos de entrenamiento, que pueden permitirle a máquinas aprender y tomar decisiones en base a esos datos, sobre todo a identificar patrones y relaciones que se ajusten a los mismos. A medida, que dichos datos van agarrando forma, el modelo del aplicativo en desarrollo se vuelve más capaz de realizar predicciones o incluso tomar decisiones. Sin alguna intervención humana directa. Así, el **Machine Learning** nos vendría ofreciendo una flexibilidad lo suficientemente notable, al adaptarse a diversos dominios, desde el reconocimiento de imágenes y voz hasta la predicción de algún análisis financiero, al igual que la contribución de recomendaciones personalizadas.

El **Machine Learning** está presente en nuestro día a día, como en el caso que cuando después de levantarnos e ir a comer o incluso antes de pararnos, tendemos a revisar rápidamente nuestro teléfono y empezamos a chequear Facebook, o alguna red social, al momento de hacer eso, ya estamos haciendo uso del Machine Learning, el cual nos estaría prediciendo que contenido es de mayor interés para uno, basándose en alguna actividad propia del pasado. [1]

Entre ese y muchos otros ejemplos, son evidentes al momento de hacer uso del **Machine Learning,** ya que este mismo cuenta con una categoría fundamental que es la del aprendizaje supervisado, donde el algoritmo se entrena con datos etiquetados, con el propósito de aprender a asociar entradas con salidas conocidas. También se destaca el aprendizaje no supervisado, que tiende a explorar datos no etiquetados, identificando patrones por sí mismo. Todo para solucionar problemas complejos, como el análisis de sentimientos en texto o la segmentación de clientes.

Motivo por el cual, no se puede negar la importancia de los **Datos** en la industria moderna, en donde podemos adentrar un término como el de **Ciencia de Datos** o **Data Scientist.** Donde el

mismo se utiliza de forma muy general, para poder indicar las prácticas y técnicas para el uso, análisis y presentación de datos. [2]

Lo que trae consigo igualmente, una serie de retos y desafíos, no sólo a nivel corporativo y empresarial, ya que, en la parte a nivel personal, también tendemos a manejar o a lidiar con altos volúmenes de información, que muchas veces no son de fácil manejo.

El caso es, que, con el uso de técnicas de aprendizaje automático, todo tipo de proyecto que sea a base de **Machine Learning,** tienen como principal objetivo el mejorar la eficiencia, la precisión y la automatización en diversas industrias.

Al sacar provecho de grandes cantidades de datos es donde entra el **Big Data**, el cual se refiere a todos aquellos datos que se caracterizan por sus altos volúmenes, velocidad y hasta variedad de información, la cual es el insumo principal para cualquier aplicación de **Machine Learning,** es ahí donde se puede llegar a generar información valiosa para la toma de decisiones estratégicas y el desarrollo de soluciones innovadoras. [3]

En las **Academias de Educación Vial,** que están de la mano del **Machine Learning** a diferencia de otras academias que no lo están, aplican esta herramienta para mejorar la enseñanza y la evaluación. **Por ejemplo,** el análisis de datos que puede identificar patrones en el desempeño de los estudiantes durante simulaciones de conducción. Lo que permite personalizar la instrucción según las necesidades individuales y así mejorar la eficacia del proceso de aprendizaje. Algunos de los sistemas utilizan algoritmos para evaluar automáticamente las habilidades de conducción, proporcionando así retroalimentación detallada a los estudiantes.

El **Machine Learning,** en sí, puede contribuir en muchos procesos de automatización, en el caso de nuestro tema abordar sobre el control y regulación del combustible de los vehículos de enseñanza, este podría contribuir con una gestión eficiente del combustible. Ya que, algoritmos de este podrían analizar datos de conducción que afecten el consumo de combustible. Permitiendo

brindar recomendaciones personalizadas a conductores en formación o a los mismos instructores, para fomentar hábitos más eficientes para el uso de estos vehículos.

Como tal, el aprendizaje automático o conocido como **Machine Learning,** se basa o apoya en varios fundamentos, entre los cuales siempre se ha de dar mayor importancia a una buena estrategia de datos. Dicha estrategia puede desenlazarse, a través de la siguiente analogía que contribuye a **una pirámide de valor de los datos** [4], la cual sigue ciertos pasos esenciales pare tener una eficiente estrategia:

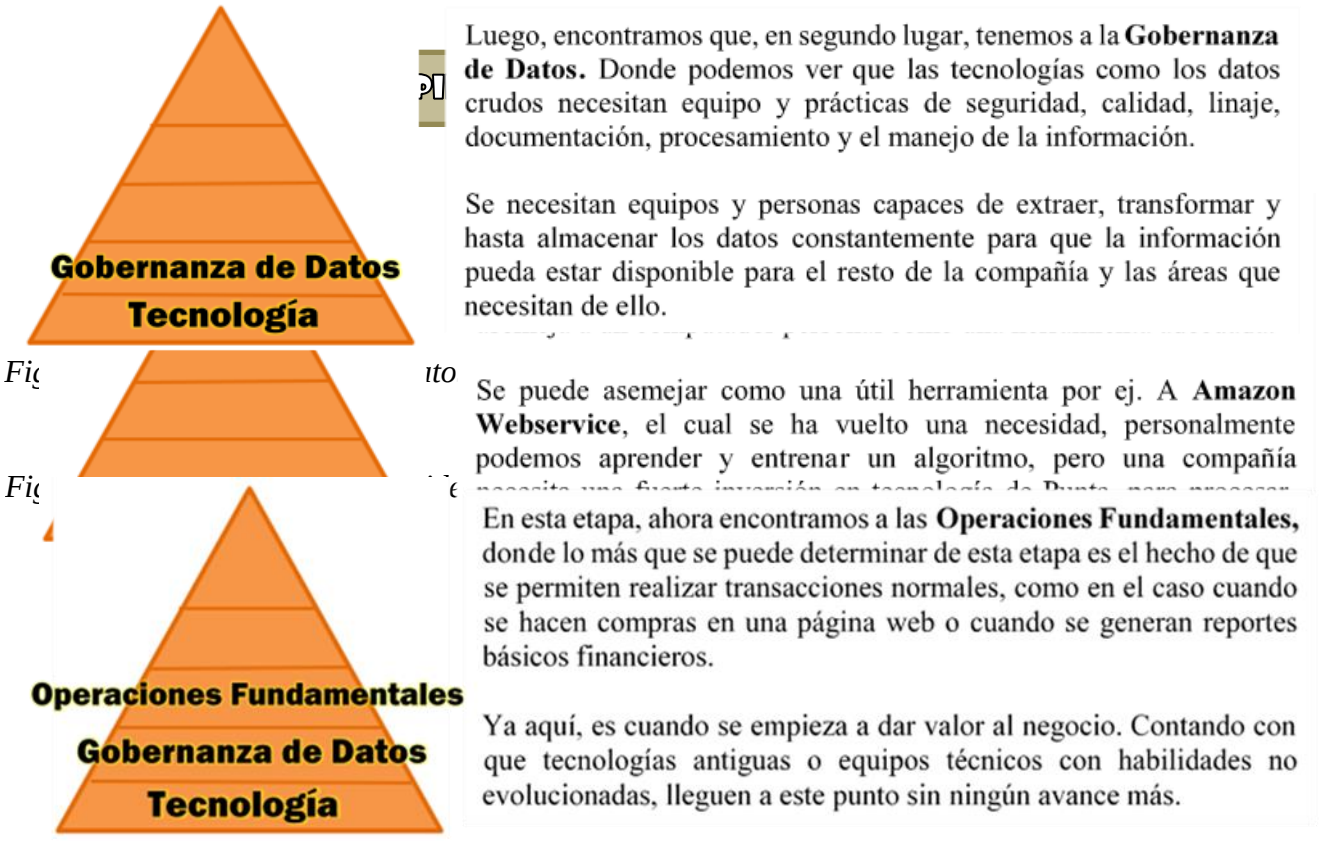


Fig. 2. Nivel 3 Operaciones Fundamentales, pirámide de valor [4]

Los pasos que siguen en la pirámide ya tienden a ser algo más exigentes y requieren de muchas más transformaciones profundas, hasta llegar prácticamente al tope.

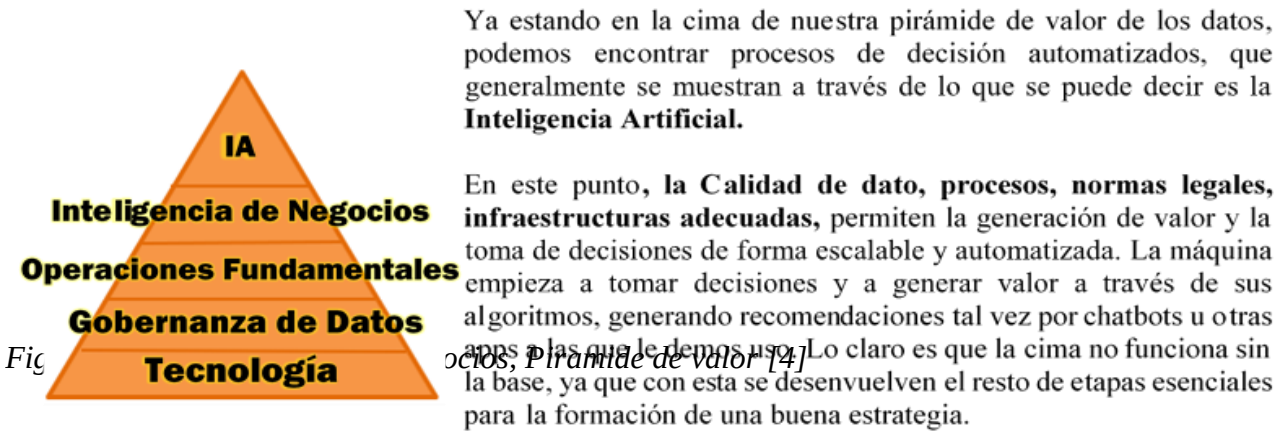
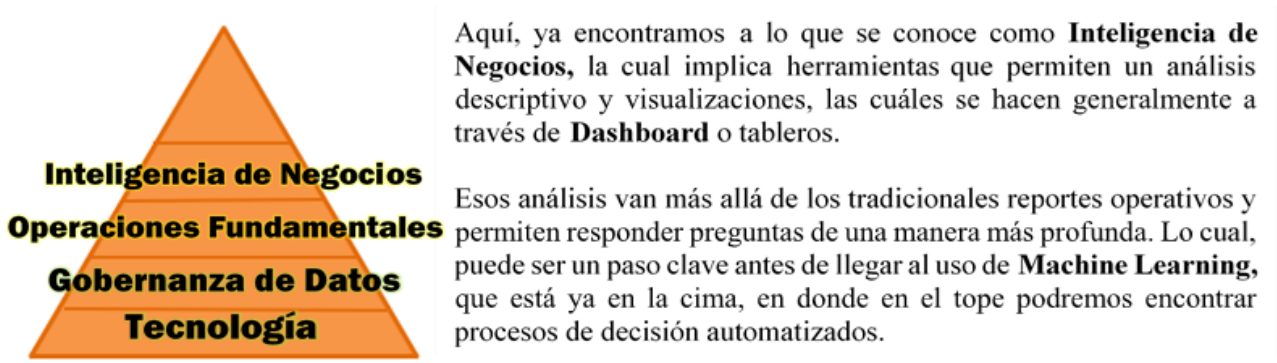


Fig. 5. Nivel 5 IA, Piramide de valor [4]

Teniendo claro todo esto, podemos ver que **Machine Learning,** es una tecnología que dependerá siempre de los datos y sin un esfuerzo en invertirle en tecnología, en procesos y demás, este no sería una realidad. El **Machine Learning** se beneficia del poder del cálculo y almacenamiento de datos para entrenar y ejecutar modelos de manera eficiente, se apoya en la retroalimentación y la iteración constante para mejorar la precisión y la capacidad predictiva de los modelos. Y todo es una combinación de estos elementos para automatizar tareas y tomar decisiones basadas en datos.



B. MACHINE LEARNING: ANÁLISIS CONTRAFACTUAL

Antes de adentrarnos en este tipo de análisis, debemos de comprender qué es lo que abarca el mismo dentro del **Machine Learning.**

Cuando hacemos referencia a un análisis contrafactual, nos referimos a algo que hubiera sido y como hubiera sido, ahí es donde entra un tipo de evaluación de como hubiera sido el resultado de un evento o momento específico, buscando determinar una acción resultante de algo que no ha sucedido. Dentro del contexto de los modelos predictivos, en este tipo de análisis lo que se implicaría sería la comparación de una predicción real con aquella predicción que se habría hecho si ciertas variables o condiciones hubieran sido distintas, lo que ayudaría a comprender la contribución de cada variable y a la vez evaluar la robustez del modelo frente a determinados cambios en los datos.

Si pudiéramos saber qué hubiera pasado en un escenario especifico o alternativo, podríamos estar claros de la estimación de buenos resultados y sobre todo resultados potenciales, en lo que se adentraría un **efecto causal.** ¿Por qué un efecto causal? Porque es ese efecto el que nos determina la comparación entre dos estados: uno verdadero en donde una acción o evento realmente sucedió y otro imaginario donde no ha ocurrido dicho evento o acción. Por ende, el contrafactual abarca un rol importante dentro del análisis, ya que ayuda a evaluar y comprender el impacto de variables especificas en las predicciones del modelo, permitiendo el establecer relaciones causales entre acciones y resultados de dichos eventos, a los que se le mide ese efecto que se da entre las dos variables al momento de compararlas y en casos en que una de esas mismas esté ausente. [5]

La importancia del análisis contrafactual radica en varios aspectos. En primer lugar, hace mejoras en la interpretabilidad de los modelos, en donde ofrece una comprensión más profunda de cómo se han de tomar las decisiones, lo que es fundamental en aplicaciones críticas en donde la transparencia es esencial. Además de que proporciona información valiosa para ajustar el o los modelos, facilita la evaluación de estos al explorar la sensibilidad de cambios en sus datos.

En el **Machine Learning**, suele destacarse más bien el hecho de que este pueda encontrar relaciones entre los datos, pues en eso abarca su capacidad de adaptación y aprendizaje automático, pero no implica el hecho de las relaciones causales, para eso se tendría que hacer un análisis completo a grandes conjuntos de datos, para así entender que variables influirían significativamente. Por eso mismo no es del todo confiable para las tomas de decisiones de negocio, por sí solo no permitiría saber la influencia que tendría una variable sobre la otra.

Dentro del **Análisis Contrafactual,** además de todo el contexto anterior es fundamental también el conocer conceptos como **Correlación** o **Causalidad**, porque la primera hace referencia a la presencia o ausencia de cualquier relación lineal entre dos variables, mientras que el otro término implica más en una relación de **Causa-Efecto,** en donde un evento o situación es causado por uno anterior. De ahí, abarca su importancia, por eso cabe aclarar que, entre ambos términos, la **Correlación** implica asociación, pero no causalidad y la **Causalidad** implica asociación, pero no **Correlación.** En resumen, el comprender estas distinciones es esencial para interpretar adecuadamente los resultados y tomar decisiones informadas. [6]

Además de eso, no se puede dejar de lado el hecho de las variables omitidas, porque de estas se pueden derivar aquellas **variables** que pueden ser **observables** o **inobservables**. Las primeras, las **observables** nos muestran a aquellas variables que no se consideraron al momento de realizar nuestro análisis, y las **inobservables,** a las que no pudieron ser observadas en los datos. Dos términos a tener muy presentes, ya que la inclusión adecuada de las mismas conlleva a aspectos cruciales para realizar buenos análisis estadísticos dentro de nuestro proyecto. [7]

Un ejemplo de ello, podría presenciarse en casos como en el estudio del rendimiento del combustible en los vehículos de las academias de conducción, en ellos se puede observar desde el tacómetro en que tanto está el combustible, y sobre todo en que puntos de las revoluciones por minuto estaríamos consumiendo más de este, o cuando no se mide el gasto que conlleva suministrarle combustible a esos vehículos para su previo uso, ya que los precios en las gasolineras podrían variar o incluso el gasto del mismo podría depender de la calidad del combustible

suministrado. La buena gestión de dichas variables es lo que nos puede llevar a conclusiones óptimas y válidas.

Como se puede ver dentro del **análisis contrafactual** en el contexto del **Machine Learning,** para la aplicación del mismo, en la mayor parte se requiere de conllevar variadas técnicas y tener presente uno que otro término para poder complementar el desarrollo de cualquier tipo de modelo que haga uso del **Machine Learning,** que además, puede constar de ciertos aspectos claves al igual que de algoritmos que pueden enriquecer el proceso de su desarrollo, aquí dentro de este trabajo se pueden determinar los siguientes aspectos y algoritmos:

Aspectos Clave del Análisis Contrafactual en Machine Learning

1. **Modelo Predictivo**

Se hace uso de un modelo de machine Learning para hacer predicciones basadas en datos históricos u observados.

2. **Escenarios Contrafactuales**

Se generan escenarios o eventos alternativos al modificar elementos o condiciones de un conjunto de datos. Lo que puede simular situaciones hipotéticas.

3. **Predicciones Bajo Condiciones Alternativas**

El modelo se aplica a estos escenarios contrafactuales para predecir cómo habrían sido los resultados si las condiciones hubieran sido diferentes.

4. **La identificación de Causalidad**

El análisis contrafactual contribuye a entender las relaciones causales al explorar cómo cambios específicos en las variables afectan los resultados, diferenciando correlación de causalidad.

5. **Interpretación de Resultados**

Se analizan las diferencias entre las predicciones reales y las contrafactuales, para poder entender la contribución de cada variable a los resultados del modelo.

6. **Aplicaciones Prácticas**

En entornos de negocios, esto podría implicar el comprender cómo ciertos cambios en estrategias afectarían a las métricas clave. Y entre otras aplicaciones como la medicina o la mecánica de autos, esto tiene un gran impacto.

7. Contrarrestar Sesgos y Mejorar la Equidad

El análisis contrafactual también se utiliza para contrarrestar sesgos en modelos de Machine Learning, lo que permite el evaluar cómo cambios en las variables podrían, mejorar la equidad y reducir discriminaciones injustas.

De esto se puede determinar que este tipo de análisis no solo mejora la manera en que interpretamos los modelos en **Machine Learning,** sino que también permite el tomar decisiones más enriquecidas de información al entender cómo los cambios en el entorno pueden influir en los resultados predichos. Lo que sería crucial en aplicaciones donde la explicabilidad y la comprensión de las decisiones del modelo han de ser esenciales.

Ahora, procedemos con los algoritmos, que tienden a aplicarse en esos modelos que hacen uso del **Machine Learning,** los cuales se reflejan en **algoritmos causales** como:

- ***Propensity Score (Puntaje de Propensión)***

Es esa herramienta estadística utilizada en estudios observacionales para abordar problemas de sesgo de selección, al estimar el efecto causal de un tratamiento. Es decir, este busca ayudar a equilibrar las características observables entre grupos tratado y no tratado. Además, se tiende a estimar típicamente mediante un **modelo de regresión logística,** donde la variable de tratamiento es la variable dependiente y las covariables son las variables independientes. [8]

Al ser estimado, ya podría hacer uso de diferentes métodos para ajustar o emparejar unidades basándose en ese puntaje. Al equilibrar las covariables observadas entre los grupos tratado y no tratado, el **Propensity score** ayudaría a reducir el sesgo de selección y a hacer que las comparaciones sean más parecidas a un experimento aleatorio.

- ***Double LASSO***

Se asemeja a una extensión del método Lasso, que se utiliza en regresiones lineales. En donde este se aplica dos veces para permitir seleccionar variables tanto en la estimación del tratamiento como en la respuesta. Ya que, lo que busca es llegar a variables más precisas que den una explicación de la variable de interés, para tal acontecimiento se seleccionan las que están más correlacionadas, luego se estima la dicha regresión por el método de mínimos cuadrados y de ahí lo que se ha de permitir es el minimizar el sesgo por variables omitidas, volviendo la estimación más precisa. [9]

Al aplicar este algoritmo, podremos llegar más pronto a soluciones ante situaciones complejas, donde no se facilita el determinar por ejemplo, que tanto consumo de combustible se difiere al utilizar los vehículos de manejo y que salga costoso en las academias de educación vial, solamente teniendo presente las estadísticas del tacómetro, caso en el que se tendrá más en cuenta otras variables, como el costo del combustible y sus variaciones por días o también el mantenimiento, con el fin de ver que tanto influye en el mal consumo y gastos inoportunos.

- ***Causal Trees***

Es un método que hace uso de árboles de decisión para modelar relaciones causales. Cada nodo del árbol representa una división basada en la variable que más contribuye a la predicción del resultado causal. Por medio de esos árboles de decisión, justo en el nodo es que se crean grupos de individuos con características similares, en donde uno de los grupos es tratado y el otro no, después es que se estima la diferencia entre el grupo de tratamiento y el grupo de Control. [10]

- ***Causal Forest***

Es una extensión o generalización del Causal Trees. Lo que quiere decir, que tiene algo de similitud en la creación de árboles de decisión, pero se diferencia en la cantidad, ya que la creación de los mismos asemejaría más bien a un bosque, cada uno centrado en modelar una relación causal diferente, que luego, promedia las predicciones de todos los árboles para obtener una estimación más concreta. Este puede encontrar efectos de tratamiento heterogéneos y a la vez mejorar los resultados del **Double Debiased Machine Learning,** sólo en el caso de si la función es altamente no lineal. [11]

De alguna manera, **Causal Forest** tiene sus ventajas, entre esas está el hecho de poder manejar tanto variables continuas como categóricas, y también el poder lidiar con datos faltantes y desequilibrados. Así, proporciona una medida de incertidumbre en las estimaciones causales. Sin embargo, puede llegar a tener sus limitaciones, debido a que requiere una cantidad suficiente de datos para obtener resultados confiables, lo que generaría un alto consumo en cuanto a recursos computacionales.

En conclusión, de todo esto se puede diferir que tanto los conceptos, como los aspectos clave del análisis contrafactual y los algoritmos, en contexto del **Machine Learning**, todos son información valiosa que puede ofrecer un gran potencial para mejorar y automatizar diversos procesos en diferentes sectores, gracias a ello es que se puede llegar a abordar problemas de sesgo de selección, estimar efectos causales y modelar relaciones complejas entre variables en entornos observacionales. Y en cuanto a la elección de algún algoritmo, ya dependerá de la naturaleza especifica de los datos y lógicamente del problema que se esté abordando.

C. EVALUACIÓN DE MODELOS DE MACHINE LEARNING

Dentro de la evaluación de modelos de **Machine Learning,** encontraremos una serie de conceptos, para aplicar en el despliegue de dichos modelos, para eso es que se determina dicha información en una serie de pasos, al estilo de una guía general para así poder desarrollar soluciones de **Machine Learning** efectivas.

1. Machine Learning Workflow

Se refiere al flujo de trabajo, aquí nosotros podemos presenciar varias etapas; en la Primera nos topamos con el entendimiento del negocio, cuál es su objetivo y cuáles las necesidades que requieran atenderse, luego se procede a lo que sería el entendimiento de los datos, con cuáles contamos y de qué tipo son, después de tener un claro entendimiento, pasamos ahora a lo que sería la preparación de los datos, y de ahí, luego al modelado, que es donde entrenamos ya nuestro

algoritmo de Machine Learning, ya una vez modelado podremos evaluarlo y finalizar la parte productiva de dicho modelo. [12]

2. Preparación de Datos

En este paso, para ejecutar correctamente un modelo de Machine Learning, se requiere de preparar el **Dataset,** que refiere a un conjunto organizado de datos, que por lo general se presentan como una instancia única, que se puede describir mediante diferentes variables. Ahí es donde podemos recurrir al uso de herramientas de programación, las cuales nos permitirán la previa visualización de datos, entre esos se han de destacar el número de observaciones y el número de variables, en caso de ser dato numérico, se haría uso de la mínima, la máxima, la desviación estándar y la media. Igualmente, se podría dar tratamiento a valores no presentes y sustituirlos por valores de promedio, si así se desea.

Luego de tener claro eso, se hace necesario el convertir las variables categóricas en datos independientes en cuanto a las variables, para eso se puede hacer uso de herramientas como **get_dummies**, debido a que esta función se ocupa de manipular los datos y el preprocesamiento en pandas (biblioteca de Python), para su proceso de transformación. En el caso de variables de tipo numérico, se puede hacer uso de la herramienta **MinMaxScaler,** ya que, con esta podremos hacer una técnica de escalado para ajustar los datos en un rango especifico y así estandarizar dichas variables numéricas. [13] [14]

3. Entrenamiento del Modelo

Ya aquí, lo que se conlleva es a un 20% del proyecto de Machine Learning, se importa lo que es el **Dataset** y se debe no olvidar el dividir los datos de entrenamiento y los de prueba. Una vez se halla corrido el algoritmo, entonces se procede a las predicciones, las cuales, buscan hacer capture de la relación que se tiene entre las variables predictivas con la variable objetivo. Además, en caso de que 2 de los algoritmos compartan un comportamiento similar, sólo se escoja el que sea de ejecución más sencilla. [15]

Además de los pasos mostrados anteriormente para la evaluación de los modelos de Machine Learning, también se tiene presente ciertas Métricas de Desempeño para el modelado y a la vez una serie de Técnicas de Validación Cruzada, que pueden complementar al mismo.

Métricas de Desempeño

Podemos encontrar las siguientes Métricas Principales:

- ***Matriz de Confusión***

Es aquella herramienta que tiene como utilidad, el evaluar justamente el rendimiento de un modelo de clasificación. Prácticamente se muestra como un tipo de tabla que muestra la cantidad de predicciones correctas e incorrectas realizadas por el modelo en cada una de las clases.

En otras palabras, lo que busca esta matriz, es resumir el desempeño que ha de tener el algoritmo en los modelos de clasificación, permitiendo visualizar el resultado de las predicciones para cada una de las observaciones. Proporcionando así, varias métricas útiles, como la precisión, la sensibilidad (**Tasa de Verdaderos Positivos**), la especificidad (**Tasa de Verdaderos Negativos**) y la Tasa de Falsos Positivos y Falsos Negativos. Métricas que ayudan a evaluar el rendimiento del modelo en términos de su capacidad para clasificar correctamente las instancias de cada clase. Lo que permitiría tomar decisiones más informadas. [16]

- ***Accuracy***

Se muestra como una derivada de la Matriz de Confusión, y es comúnmente utilizada para evaluar el rendimiento de un modelo de Machine Learning. Representa la proporción de predicciones correctas realizadas por el modelo en relación al total de predicciones realizadas

Esta determina el número de predicciones correctas divididas en el total de predicciones. Haciendo uso de la siguiente fórmula se puede calcular su exactitud:

$$\frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{TP})}$$

Del resultado que sea determinado, nos podrá brindar el porcentaje de efectividad que tendrá en las predicciones. Sin embargo, una desventaja de esta métrica podría verse reflejada en el hecho de que oculta la clasificación de las observaciones, lo que podría ocasionar que se genere un porcentaje alto en etiquetas o tazas negativas, aun así, eso se podría solucionar con el uso de otras métricas que se pueden determinar con la matriz de confusión. [17]

- ***Precisión y Recall***

Refieren a dos métricas utilizadas en la evaluación de algoritmos de clasificación y recuperación de información. Ambas métricas, son importantes y están relacionadas con los objetivos del problema. Si lo que se busca es minimizar los falsos positivos (clasificar algo como positivo cuando en realidad es negativo), se debe tener presente la **Precisión.** Por otro lado, si lo que se quiere es minimizar los falsos negativos (clasificar algo como negativo cuando en realidad es positivo), es importante el considerar a el **Recall.**

La **Precisión,** se calcula dividiendo el número de verdaderos positivos entre el total de la suma de verdaderos positivos y falsos positivos. En cuanto al **Recall,** este ya vendría siendo el resultado de dividir el número de verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos. [18]

- ***Specificity y F1 Score***

Los dos son métricas comunes utilizadas en la evaluación de modelos de clasificación. La primera mide la capacidad de un modelo para identificar correctamente los casos negativos. Es decir, indica qué tan bueno es el modelo para predecir los casos negativos cuando realmente son negativos. Mientras que la **F1 Score,** se ocupa de calcular como, la media armónica entre la precisión y el Recall.

El **Specificity,** en otras palabras, se muestra como el caso contrario del Recall, lo que quiere decir, que, su resultado será la división de los verdaderos negativos y la suma de los verdaderos negativos y los falsos positivos. Mientras que, con el **F1 Score,** lo que se ha de abarcar es la

combinación de la Precisión y el Recall, y su resultado será la división de la Precisión por el Recall, en donde a dicho resultado se le multiplica por 2. [19]

- ***Curva ROC y AUC***

La **Curva ROC** hace parte de una representación gráfica de la sensibilidad y especificidad de un modelo de clasificación de tipo binario, esta busca mostrar el cómo cambia el equilibrio entre los mismos a medida que se ajusta el umbral de probabilidades dentro de la clasificación. Mientras que el **AUC,** es ese valor numérico que resume la **Curva ROC** en un solo número. El uso de estos es muy común para dicha evaluación de modelos, ya que proporcionan una medida robusta y visualmente interpretable del rendimiento del modelo.

Tienen muy presente el área bajo de la curva, entonces por eso mismo entre mayor sea el área mejor será el comportamiento del modelo. Al igual que en el caso de que la curva disminuya también lo hará el comportamiento del modelo.

- ***Curva de Precisión-Recall***

Muestra la relación entre la precisión y el Recall (conocido también como sensibilidad) a medida que se va ajustando el umbral de clasificación. Al ajustar el mismo, se puede obtener diferentes pares de valores de **Precisión** y **Recall.** Entonces es ahí en la curva donde se muestra esos puntos como un gráfico. De ahí, se empieza a evaluar el rendimiento del modelo en función del equilibrio deseado entre **Precisión** y **Recall.**

Es especialmente útil cuando los datos están desequilibrados, es decir, cuando existe una gran diferencia en el número de instancias positivas y negativas. Ahí se grafica en el eje **Y** la **Precisión** y el **Recall** en el eje de la **X.** [20]

Técnicas de Validación Cruzada

Dentro de estas técnicas, se nos permite el realizar un muestreo entre los datos de entrenamiento y los datos de validación, los cuales nos permiten el poder mirar que tan efectiva es la capacidad de predicción correcta, al igual que el poder estimar el comportamiento ante nuevos datos que no se tuvieron en cuenta en el modelo. [21] Entre esas Técnicas se pueden destacar algunas como:

- ***K-Fold Cross Validation***

Se basa en una técnica para evaluar y seleccionar modelos de manera más robusta. Aquí se tiende es a dividir el conjunto de datos en **K** particiones o **"folds"** de tamaño similar. Luego, se entrena y evalúa el modelo **k** veces, cada vez haciendo uso de una partición diferente como conjunto de prueba y las demás particiones como conjunto de entrenamiento.

- ***Stratified K-Fold Cross Validation***

Es usada en el **Machine Learning,** para evaluar la eficacia de un modelo de clasificación o regresión. A diferencia de **K-Fold Cross Validation** tradicional, donde lo que sucede es que se divide el conjunto de datos en K partes iguales, en la **Stratified K-Fold Cross Validation** se asegura más bien de que cada parte tenga una proporción similar de muestras de cada clase.

Al estratificar los datos en cada fold, se garantiza que cada uno de ellos tenga una representación proporcional de todas las clases, lo que ayudaría a evitar sesgos en la evaluación del modelo.

- ***Leave-P-Out Cross Validation (LpO)***

Esta técnica, en lugar de dividir los datos en conjuntos de entrenamiento y prueba de tamaño fijo, lo que hace es dejar "P" observaciones fuera del conjunto de entrenamiento y las utiliza como un conjunto de prueba. Proceso que se repite para todas las combinaciones posibles de dejar "P" muestras fuera.

Es útil cuando se tienen conjuntos de datos pequeños y se desea obtener una estimación más precisa del rendimiento del modelo. Al probarlo en diferentes conjuntos de prueba, se reducen los sesgos y la varianza asociados con una única división de entrenamiento/prueba.

- ***Leave-One-Out Validation (LOO)***

Con esta técnica lo que se hace es entrenar y evaluar el modelo, utilizando todos los datos excepto uno. Por cada iteración, se selecciona un único punto de datos como conjunto de prueba y se entrena el modelo con el resto de los datos. Luego, se evalúa el rendimiento del modelo utilizando el punto de datos excluido. Lo que lo hace algo beneficioso cuando se trata de un conjunto de datos pequeños, sin embargo, si el conjunto de datos es grande, puede resultar costoso computacionalmente.

Cada uno de estos pasos, como las métricas de desempeño y las técnicas de validación cruzada, son algo fundamental, si se quiere medir la calidad y el rendimiento de los modelos. En conclusión, el evaluar modelos de Machine Learning a través de estos procesos, promueve y garantiza la fiabilidad y la eficacia de los modelos en diferentes escenarios y conjuntos de datos.

D. MACHINE LEARNING APRENDIZAJE SUPERVISADO

Para entender cómo influye este tipo de aprendizaje en el Machine Learning debemos de comprender en que se basa este, por eso vamos directo al **Aprendizaje Supervisado.**

El **Aprendizaje Supervisado,** es una rama del Machine Learning, por eso es que se le puede llegar a conocer como Machine Learning Supervisado, y se muestra como una subcategoría del Machine Learning y La Inteligencia Artificial. Definiéndose por su uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifican datos o anticipan resultados con precisión.

A medida que se van introduciendo estos datos en el modelo, se van ajustando sus ponderaciones hasta que el modelo se adapte correctamente, proceso que hace parte de la validación cruzada. Es ahí que el **Aprendizaje Supervisado,** les permite a organizaciones resolver

una amplia variedad de problemas de la realidad a modo escala. Debido a que su objetivo principal es que los modelos o el modelo aprenda a hacer predicciones precisas sobre nuevos datos no vistos previamente. [22]

En el caso de nuestro proyecto, el paso a proseguir es el realizar las primeras predicciones y luego de eso, adherir una clasificación estructurada en partes que podrían ir variando según el comportamiento de los escenarios que estén en proceso de estudio.

Además, en este tipo de aprendizaje, también podemos encontrar algoritmos que ayudan a contribuir en patrones y relaciones entre las características de entrada y salida deseadas. Los cuales pueden ser lineales, como la regresión lineal, o no lineales, como las redes neuronales. Pero en el punto del entrenamiento, es donde podremos ver que cada parámetro se ajuste con el fin de poder minimizar la diferencia entre las predicciones y las respuestas reales en el conjunto de datos de entrenamiento.

En base a esos algoritmos, podemos denotar el uso de varios de ellos y sobre todo ciertas técnicas de cálculo, que, en este caso, se pueden implementar con la herramienta de Python. De esto se pueden derivar los siguientes conceptos dentro de los algoritmos aplicados:

- ***Regresión Lineal***

Se utiliza para identificar la relación entre una variable dependiente y otra independiente, además, esta toma provecho para realizar o estipular predicciones sobre futuros resultados. Cuando solo se presenta una variable independiente y una variable dependiente, a esto se le conoce o determina como una **regresión lineal simple.** Pero a medida que va aumentando el número de variables independientes, se hablaría entonces de una **regresión lineal múltiple.** En todo caso, para cada tipo de regresión lineal, se debe de trazar la línea que mejor se ajuste, y de ahí se podrán hacer cálculos mediante el método de mínimos cuadrados. Sin embargo, a diferencia de otros modelos de regresión, la línea podría ser recta en casos en el que trace en un grafo. Entonces para que las variables estén de lo más ordenadas, se puede utilizar el algoritmo de mínimos cuadrados ordinarios, el cual, ajustaría todos los datos. [23]

- ***Regresión Logística***

A diferencia de la regresión lineal, esta se utiliza cuando las variables dependientes son continuas, se le elije cuando la variable dependiente es categórica; lo que quiere decir, cuando esta tiene salidas binarias, tales como “Verdadero” y “Falso” o “Sí” y “No”. Ambos modelos de regresión tratan de comprender las relaciones entre las entradas de datos, sin embargo, la regresión logística se utiliza fundamentalmente para resolver problemas de clasificación binaria, como la identificación de correo no deseado o spam.

Además, al ser una regresión que se aplica a variables de clasificación binaria, es de suma importancia el tener presente que tanto palabras como textos deberán ser convertidos a números. [24]

- ***Árboles de Decisión***

Son un algoritmo de aprendizaje supervisado no paramétrico, que es utilizado para resolver problemas tanto de clasificación como de regresión. Y se muestra como una estructura de tipo jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Cada nodo interno representa una pregunta sobre una característica, mientras que las ramas representan las posibles respuestas a esa pregunta. Y los nodos hoja contienen las etiquetas de clasificación o los valores de regresión final.

De esto se puede determinar que estos árboles son fáciles de interpretar y visualizar, lo que los hace útiles para tomar decisiones explicables. Permiten el clasificar segmentos de acuerdo con los datos. [25]

- ***Redes Neuronales***

Son un tipo de algoritmo de aprendizaje automático inspirado en el funcionamiento del cerebro humano. Están compuestas por capas de neuronas artificiales interconectadas, que procesan y transmiten información a través de conexiones ponderadas.

Se utilizan principalmente para los algoritmos de Deep Learning, ya que procesan los datos de entrenamiento imitando la interconectividad del cerebro humano a través de capas de nodos. En donde cada uno está formado por entradas, ponderaciones, un sesgo o umbral y una salida. Si el valor de salida se excede, entonces ahí es cuando se activa el nodo, suministrando datos a la siguiente capa de la red. De esta forma, es que estas redes neuronales aprenden a correlacionar a través del aprendizaje supervisado, haciendo ajustes en base de la función de pérdida a través de procesos como su gradiente descendiente. En casos de que la función de coste sea igual o se acerque a 0, podremos confiar en que la precisión de este modelo podrá obtener la respuesta correcta.

Además, en este tipo de redes, se derivan 3 partes específicas para poder tratar tareas determinadas en la amplia variedad de modelos, esas son: el **input layer,** que permite su entrenamiento, el **hidden layer,** que promueve la activación de las mismas y el **output layer,** que será el que brinde la predicción o clasificación.

Con esto podemos decir que, las redes neuronales pueden ser muy efectivas en tareas complejas de reconocimiento y clasificación de patrones, sin embargo, también tienden a requerir de grandes cantidades de datos de entrenamiento y el poder computacional para alcanzar su máximo potencial. [26]

De todo esto se puede concluir que, el Machine Learning bajo un aprendizaje supervisado se vuelve en un campo poderoso y ampliamente utilizado, Permitiendo automatizar tareas complejas y haciendo predicciones precisas en una variedad de aplicaciones, sin embargo, es fundamental el tener presente que el rendimiento del modelo supervisado dependerá en gran medida de la calidad y la cantidad de datos de entrenamiento disponibles.

E. FUNDAMENTOS APLICADOS DE MACHINE LEARNING

Python es un lenguaje de programación ampliamente utilizado en el campo de Machine Learning debido a su gran comunidad de desarrolladores y a las herramientas y librerías específicas desarrolladas para este propósito. Algunas de las librerías clave para el curso incluyen:

NumPy: Esta librería, desarrollada desde 2005, tiene como objetivo habilitar la computación numérica en Python [27], lo que la hace fundamental para el análisis de datos y la implementación de algoritmos de machine Learning.

Pandas: Desarrollada desde 2008, Pandas es fundamental para la realización de análisis de datos práctico en el mundo real. Proporciona estructuras de datos flexibles y herramientas de manipulación de datos que son esenciales en el contexto de machine Learning.

Scikit-Learn: Esta librería, que existe desde 2007, ofrece herramientas simples y eficientes para el análisis de datos predictivos [27]. Se integra muy bien con NumPy, Pandas, Matplotlib y otras librerías, lo que la hace una opción poderosa para la implementación de algoritmos de machine Learning.

1) Ambientes de Desarrollo e Interfaces:

Además de las librerías, es importante familiarizarse con los ambientes de desarrollo y las interfaces que facilitan el trabajo en Python para machine Learning. Un ejemplo es:

- **Jupyter Notebooks:** Este ambiente de desarrollo permite trabajar en el navegador y combinar texto, ecuaciones, gráficos y código. Es ideal para propósitos educativos, documentar procesos y compartir información. Es ampliamente utilizado y cuenta con soporte directo en herramientas de grandes empresas como Google y Amazon [27].

2) Preprocesamiento de Datos y Análisis Exploratorio:

El preprocesamiento de datos es un paso importante en cualquier proyecto de Machine Learning. El objetivo del preprocesamiento es preparar los datos para el entrenamiento del modelo.

Algunas técnicas de preprocesamiento comunes incluyen:

1. **Eliminación de valores atípicos:** los valores atípicos son datos que se desvían significativamente del resto de los datos. Estos valores pueden afectar negativamente el desempeño del modelo, por lo que es importante eliminarlos.
2. **Limpieza de datos faltantes:** los datos faltantes pueden ocurrir por varias razones, como errores de entrada o problemas de medición. Es importante abordar los datos faltantes antes de entrenar el modelo.
3. **Transformación de variables:** algunas variables pueden necesitar ser transformadas para que el modelo las entienda mejor. Por ejemplo, las variables categóricas pueden ser transformadas a variables numéricas.

En el análisis exploratorio de datos (EDA), se propone estudiar la variable objetivo mediante estadísticas descriptivas y visualización gráfica [28]. Herramientas como sns.lmplot() y matrices de correlación son empleadas para analizar las relaciones entre variables.

3) Estimación del Modelo y Ejercicio de Clasificación:

La estimación de modelos es el proceso de entrenar un modelo de Machine Learning a partir de datos. Existen muchos algoritmos diferentes de estimación de modelos disponibles, cada uno con sus propias ventajas y desventajas. La distinción entre set de entrenamiento y test se destaca en la estimación del modelo, donde se introducen las métricas principales para regresión: RMSE y R2.

En el ejercicio de clasificación, se aborda el proceso focalizado en datos no reportados o mal reportados. Se revisan variables numéricas, se detectan y manejan valores no reportados, y se aplican técnicas como "One hot encoding" para variables categóricas. La base de datos se divide en conjuntos de entrenamiento y test.

4) Escalado de Variables Numéricas y Regresión Logística:

El escalado de variables numéricas es esencial. Se instancia la regresión logística y se evalúa el modelo utilizando la matriz de confusión como herramienta clave para ejercicios de clasificación.

5) Ciclo de Vida de un Proyecto y Tipos de Aprendizaje:

Se introduce el ciclo de vida de un proyecto de machine Learning, desde la definición del problema hasta la puesta en producción. Se distinguen dos tipos de aprendizaje:

- **Supervisado:** (Fig. 6.), donde se tiene información sobre el resultado esperado, En el sentido en que se tiene una observación o una realidad sobre el resultado que esperaría de una predicción [29].
- **No supervisado:** (Fig. 7.), también conocidas como técnicas de clustering o agrupaciones. Son más útiles cuando no se tiene variables asociadas y permiten agrupar, de acuerdo a diferentes categorías o técnicas de semejanza las observaciones que se tiene [29], es decir el modelo se entrena con datos que no contienen la variable objetico.

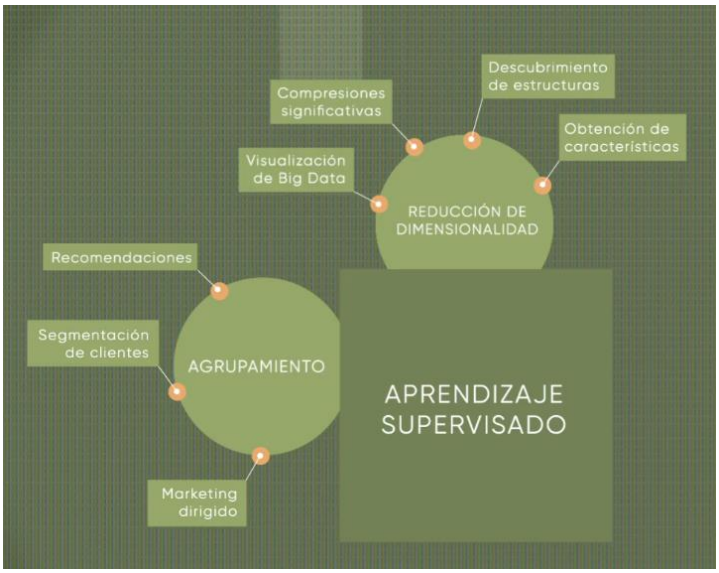


Fig. 6. Aprendizaje supervisado

Fig. 7. Aprendizaje no supervisado

F. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) es un campo de la informática que se ocupa de la creación de sistemas inteligentes, capaces de aprender y adaptarse. Se distinguen dos enfoques: la Inteligencia Artificial General, que imita las capacidades cognitivas humanas, y la Inteligencia Artificial Narrow, que se centra en características o capacidades específicas, como el reconocimiento de texto, interpretación de audio e imágenes, clasificación, entre otros.

Principios de Machine Learning según la Universidad de Wharton

- **Generalización:** Tenemos que buscar que no sean exactos, estos modelos matemáticos nos acercan a cómo funciona la realidad [30].
- **No free lunch:** No hay un algoritmo de ML que sea el mejor para todas las tareas. El algoritmo adecuado depende del problema específico que se desea resolver.
- **La navaja de Ockham – principio de parsimonia:** Es mejor hacer los modelos y algoritmos lo más simples posible.

- **Mas datos es mejor que algoritmos más complejos:** La abundancia de datos, como en el caso de Google, contribuye a la potencia de la inteligencia artificial.
- **Validación Cruzada:** Es una técnica utilizada para evaluar el rendimiento de un modelo de machine Learning.
- **Diversidad de algoritmos:** No se debe limitar a utilizar un solo algoritmo. Es importante buscar el algoritmo adecuado para cada tarea y que sea uno de los mejores para alcanzar la métrica deseada.

6) Diferencias entre AI, ML, DL y Data Science

- **Inteligencia artificial (AI):** Campo de la informática que se ocupa de la creación de sistemas inteligentes, capaces de aprender y adaptarse.
- **Machine Learning (ML):** Rama de la IA que se ocupa del desarrollo de algoritmos que aprenden de los datos.
- **Deep Learning (DL):** Una subcategoría de ML que utiliza redes neuronales artificiales para aprender de los datos.
- **Data Science:** Es el proceso de combinar estadística y programación para trabajar con grandes volúmenes de datos y llegar a generalizaciones o modelados de fenómenos [31].
- **Big data:** Hace referencia a tres características, una de ellas es la velocidad, que implica la generación de datos en tiempo real o a altas velocidades. Esto permite conectarse a la información, vaciarla en una base de datos y utilizarla posteriormente. La segunda característica es el Volumen, donde se manejan gran cantidad de datos. Y por último se encuentra la Variabilidad, la cual hace referencia a datos estructurados, semiestructurados y no estructurados. Los **datos estructurados** son aquellos que se comportan como si fueran una tabla de Excel muy grande. En la tabla Excel se tienen los nombres de las columnas, ejemplos o filas, el tipo de dato que se encuentra en cada columna. Por otro lado, están los datos **no estructurados**, como, por ejemplo, videos, audios, imágenes. Y los **Datos semiestructurados,** es una información que está más organizada.

7) Deep Learning

El interés en Deep Learning se vincula al auge de la inteligencia artificial. Estas técnicas han mejorado la precisión de modelos, permitiendo la identificación, reconocimiento y descripción de interacciones humanas. La abundancia de datos conectados a internet facilita la creación de redes neuronales más profundas, aunque el procesamiento de grandes cantidades de datos requiere capacidades computacionales considerables.

Las redes neuronales son un tipo de algoritmo de ML que se inspira en el funcionamiento del cerebro humano. Las redes neuronales están formadas por una serie de neuronas que se conectan entre sí. Cada neurona recibe entradas de otras neuronas y produce una salida que se envía a otras neuronas. Además, se pueden utilizar para realizar una amplia gama de tareas, como la clasificación, la regresión y la detección de anomalías.

La arquitectura de redes neuronales

hace referencia no solo al número de capas si no al número de neuronas que tienen dentro de ellas [32]. Es decir, se refiere a la estructura de las neuronas y las conexiones entre ellas. Las redes neuronales pueden tener diferentes arquitecturas, como redes neuronales feedforward, redes neuronales convolucionales y redes neuronales recurrentes.

Por un lado, la Red feed forward se caracteriza por tener capas conectadas entre sí, que a su vez tienen neuronas, estas neuronas se conectan con las de la siguiente capa. Pasando información desde un extremo de la red que puede ser la capa de entrada, hasta el otro extremo que es la capa de salida. Por otro lado, el Perceptrón es una de las arquitecturas más básicas de las redes neuronales. consta de dos celdas de entrada y una de salida. Y las redes Convolucionales se utiliza para el procesamiento de imágenes.

Teachable Machine es una de las herramientas que se usa para entrenar un modelo de machine Learning, esta permite modelar tareas de visión computacional, como el reconocimiento de objetos y el seguimiento de movimientos.

Dimensionalidad de los datos

La dimensionalidad de los datos es un concepto importante en el campo de la inteligencia artificial, ya que nos permite comprender la estructura y la forma en que se presentan los datos de entrada y salida.

- **Dimensión cero:** Se refiere a datos escalares, es decir, un único dato individual.
- **Dimensión 1:** Corresponde a los vectores, que son conjuntos de varios datos escalares organizados en forma de listas o diccionarios.
- **Datos en 2D:** Estos datos son comúnmente importados o exportados desde Excel y se presentan en forma de matrices o tablas. Se les llama 2D porque están dispuestos en filas y columnas.
- **Datos en 3D:** Representan datos que tienen una tercera dimensión, lo que puede ser útil en aplicaciones como imágenes en blanco y negro, donde se agrega la información de profundidad.
- **Datos en 4D:** Se refiere a situaciones en las que entran en juego las imágenes a color, ya que agregan una cuarta dimensión para representar las distintas matrices de color (rojo, verde y azul) que componen la imagen.
- **Tensores:** Este concepto se utiliza cuando se trabaja con más de 3 dimensiones, lo que resulta fundamental en el procesamiento de datos complejos, como en el caso de imágenes tridimensionales o vídeos.

G. INTRODUCCIÓN A LA ÉTICA EN LA INTELIGENCIA ARTIFICIAL

Para abordar este tema que poco se le presta atención en la tecnología, se trae una frase muy conocida por el escritor Gerd Leonhard y mencionada en el primer curso de este módulo "La tecnología no tiene ética, pero la humanidad depende de ella" [33], sugiere que, por sí misma, la tecnología no posee un conjunto de principios éticos o morales. La tecnología es una herramienta, un medio para lograr objetivos, y su desarrollo no está intrínsecamente vinculado a consideraciones éticas. Sin embargo, la segunda parte de la frase destaca que, a pesar de la carencia inherente de ética en la tecnología, la humanidad se ha vuelto profundamente dependiente de ella.

El uso de la tecnología plantea una serie de problemáticas que deben ser abordadas para garantizar un uso justo y ético. Entre estas problemáticas se encuentran los sesgos y prejuicios, los derechos de autor y la transparencia en el desarrollo y uso de modelos de IA:

Sesgos y prejuicios

Los sesgos y prejuicios son inherentes e intrínsecamente a la experiencia humana. La inteligencia artificial no tiene la capacidad humana de darse cuenta de sus prejuicios y crecer a partir de ellos. Esto puede conducir a resultados sesgados en los modelos de IA, que pueden perpetuar la discriminación y la desigualdad.

Derechos de Autor:

Debate sobre la creación de contenido a partir de modelos de inteligencia artificial. Existe preocupación por posibles violaciones de los derechos de autor de escritores, artistas y productores de contenido.

Transparencia en el desarrollo y uso de modelos de IA:

Es esencial para justificar las decisiones en tecnología y proteger los derechos de los usuarios y la sociedad en general.

De igual manera se han buscado soluciones a estas problemáticas, incluyendo las siguientes, cabe destacar que no hay una solución que los mitigue, pero trabajando en conjunto se logra un buen resultado, por ejemplo:

Las investigaciones sobre sesgos en IA están en constante evolución, por lo que es importante mantenerse al día con las últimas investigaciones para comprender mejor cómo se pueden mitigar los sesgos.

Es importante desarrollar una definición clara de lo que significa la creación de contenido a partir de modelos de IA para que se puedan determinar los derechos de autor de manera justa.

Publicar información sobre los algoritmos utilizados para desarrollar los modelos para que los usuarios puedan comprender cómo funcionan los modelos.

En el desarrollo de un modelo predictivo de consumo de combustible, es imperativo abordar cuestiones éticas clave. Se debe prestar especial atención a la identificación y mitigación de posibles sesgos y prejuicios en los datos utilizados para entrenar el modelo. Además, se deben abordar de manera ética las preocupaciones relacionadas con derechos de autor al crear contenido a partir del modelo. La transparencia en el proceso, incluida la divulgación de la metodología y los

algoritmos utilizados, es esencial para establecer una base ética sólida y fomentar la confianza de los usuarios y la sociedad en general.

En el contexto del proyecto de implementación de un modelo predictivo de consumo de combustible para la escuela de conducción, los datos disponibles, que incluyen detallados recibos de máquina por la compra de combustible y registros en libros, desempeñan un papel crucial en abordar aspectos éticos relacionados con la inteligencia artificial. Desde la perspectiva de sesgos y prejuicios, el análisis de estos datos permite una exploración exhaustiva en busca de posibles distorsiones en los patrones de consumo. Al examinar variables como el nombre del instructor, el tipo de vehículo y el establecimiento de compra, se busca identificar y mitigar sesgos no deseados.

En términos de derechos de autor, la información detallada en los recibos y registros en libros establece de manera clara la propiedad de los datos y cualquier contenido generado, evitando ambigüedades sobre la autoría y derechos de autor asociados. La solicitud de facturas electrónicas al comprar combustible demuestra un compromiso adicional con el respeto de los derechos de autor, y la colaboración con expertos legales puede asegurar que estos procesos estén en conformidad con las leyes pertinentes.

En lo que respecta a la transparencia en el desarrollo del modelo, la documentación detallada presente en los recibos y registros en libros sirve como base para explicar claramente cómo se seleccionaron y procesaron los datos, qué algoritmos se utilizaron y cómo se entrenó el modelo. Informar a los instructores sobre la implementación del modelo y su contribución a la eficiencia operativa fomenta la transparencia, y la solicitud de facturas electrónicas demuestra una práctica transparente y verificable.

H. INNOVACIÓN TECNOLÓGICA CON INTELIGENCIA ARTIFICIAL

En la actualidad, el mundo experimenta una transformación notable, dejando atrás métodos manuales para dar paso a una era de automatización. Este cambio no solo ha redefinido la manera en que llevamos a cabo tareas cotidianas, sino que también ha creado conexiones que antes parecían inverosímiles, ahora al alcance de unos simples clics. La disponibilidad de cantidades

colosales de información en dispositivos móviles plantea interrogantes cruciales sobre el futuro de los negocios, las economías globales y las políticas públicas.

En este contexto de evolución tecnológica acelerada, surge la pregunta inevitable: ¿Cómo podemos enfrentar y competir con la innovación tecnológica que está tocando a nuestras puertas? En este contexto, exploraremos las tendencias y desafíos que acompañan a la revolución tecnológica actual y cómo podemos adaptarnos para aprovechar las oportunidades que ofrece la inteligencia artificial.

En las últimas décadas, se ha logrado emular procesos y sistemas naturales a escalas más pequeñas mediante diversas técnicas. Los algoritmos genéticos, inspirados en la evolución, siguen la premisa de la supervivencia del más apto. La Lógica Difusa replica la aproximación lingüística del sistema parasimpático para tomar decisiones en situaciones donde la precisión es limitada pero la conciencia es suficiente. Por su parte, los Algoritmos Virales imitan el comportamiento de virus masivos para explorar universos de alta complejidad y extensión [34].

La innovación, por su parte, tiene como **objetivo resolver problemas** aparentemente sin solución. A través de la exploración de diversas áreas y la conexión de conceptos contra patrones, es posible destilar ideas que se acerquen a propuestas innovadoras.

Dentro de la entidad se ha identificado un problema a evaluar, en primer lugar, se destacan las reformas implementadas por el gobierno actual, las cuales incluyen un aumento gradual en el precio del combustible. Este cambio ha llevado a la necesidad de llevar un registro preciso del gasto y consumo de combustible en los recorridos semanales. En el ámbito interno, se observa que existen discrepancias en el consumo de combustible entre los tres instructores. En particular, se identifica que uno de ellos está consumiendo más combustible que los demás, lo cual resulta ser una situación irregular.

Esta disparidad en el consumo de combustible plantea la urgencia de desarrollar estrategias para optimizar el uso de estos recursos dentro de la institución. La implementación de medidas eficientes no solo permitirá economizar recursos, sino que también contribuirá a prevenir posibles casos de fraude dentro de la organización.

La identificación de oportunidades para innovar implica abordar problemas existentes o mejorar procesos eficientes pero susceptibles a optimización. En este contexto, la recopilación, uso y manejo de datos en el diseño y desarrollo de soluciones con inteligencia artificial son cruciales. La obtención de datos se puede realizar mediante técnicas como el **web scrapping**, una herramienta automatizada que extrae información relevante de internet, permitiendo una investigación efectiva sobre propuestas de innovación.

8) Tipos de Inteligencia Artificial:

Dentro del ámbito de la inteligencia artificial, se distinguen dos tipos principales: la IA General, que busca emular sistemas tan complejos como el cerebro humano en su totalidad, y la IA Especializada, que se enfoca en tareas específicas. La IA Especializada incluye ramas populares como el aprendizaje automático o Machine Learning, algoritmos evolutivos y la lógica difusa.

Los algoritmos evolutivos optimizan herramientas mediante simulaciones, determinando condiciones óptimas para sobrevivir en entornos particulares. La lógica difusa, por otro lado, emula problemas con parametrización lingüística, brindando resultados estimados robustos en situaciones donde la precisión numérica no es imperativa.

Profundizando en el Aprendizaje Automático, se distinguen tres tipos: Aprendizaje, Supervisado, No Supervisado y por Refuerzo. Cada uno representa enfoques diferentes para representar y procesar información, abriendo posibilidades diversas en la aplicación de la inteligencia artificial. La diferencia es como los datos están representando la información, cuando se tiene parametrizada la información que interesan, es supervisado, por ejemplo, si se quisiera hacer un modelo que nos diga si una opinión es positiva o negativa, la base de datos ya debería tener alguna clasificación con ejemplos de, cual es una opinión buena y cual es una opinión mala [35]. En cambio, el aprendizaje no supervisado no tiene ninguna parametrización de este tipo, solo se tienen datos y no se sabe del todo como se deberían agrupar, entonces el algoritmo genera un sistema para separar diferentes grupos de datos encontrando patrones entre las diferentes variables, con base en eso el aprendizaje no supervisado devolverá tantos grupos como haya encontrado.

II. DESARROLLO E IMPLEMENTACIÓN

En la Actualidad, sería imposible imaginarse una vida sin un medio de transporte, que se ha convertido en un elemento fundamental en el día a día. Por eso mismo, el automóvil se ha considerado uno de los principales logros tecnológicos dentro de la historia de la humanidad. Es desde este punto que la industria del automóvil ha experimentado profundas y significativas transformaciones en los últimos años. De ahí nace la necesidad del ser humano, de estar cada vez más incorporado en los avances tecnológicos en la sociedad, ya que no solo una persona, sino varias son las personas que tienen la necesidad de poder trasladarse de un lado a otro, es por eso que, ha ido aumentando la necedad de poder valerse por sí mismo en la complejidad del tráfico, que cada vez más con el crecimiento constante de la población se torna más complicado, debido a la presencia de más vehículos en las carreteras. Motivo por el cual las academias de conducción o educación vial se han vuelto toda una tendencia, ya que ofrecen programas que enseñan a conductores no solo habilidades básicas de manejo, sino también estrategias para enfrentar situaciones de tráfico complejas y emergentes, por eso se tiende a usar mucho los vehículos de enseñanza constantemente, y de ahí puede variar su desgaste y el consumo de combustible.

Debido a esto, este proyecto se enfocará en mejorar los costos ocasionados en cuánto al consumo de combustible de los vehículos de enseñanza para la obtención de las licencias de conducir. Ya que no hay un área específica, que se encargue de hacer una buena gestión y control de la información que entra y sale con respecto al gasto del combustible, los datos son llevados rudimentariamente a través de libros físicos, pero no hay una trazabilidad como tal de dicho proceso. Por tanto, esta iniciativa se propone utilizar Técnicas de Machine Learning para abordar el problema, específicamente a través de la aplicación de la regresión lineal múltiple utilizando datos estructurados. De este modo, surge el siguiente interrogante: ¿Cómo podemos predecir de manera eficiente el consumo de combustible basándonos en datos históricos y variables relevantes?

A. WORKFLOW

1) ***Entendimiento del negocio***

La "**Escuela de Automovilismo Continental del Norte**" es una empresa ubicada en Popayán-Cauca, la cual se especializa en la expedición de licencias de conducir para carros y motos. También ofrece cursos de enseñanza para la obtención de dichas licencias. En términos de gestión de datos, la entidad opera de manera rudimentaria, registrando la información en libros físicos sin una trazabilidad clara. La responsabilidad de manejar toda la información recae en una persona, quien también desempeña diversas tareas administrativas, como la recepción de estudiantes y la gestión de pagos.

Por otro lado, el objetivo principal de este proyecto es optimizar los costos asociados al consumo de combustible de la flota de vehículos utilizados en las clases. Debido a esto, surge la necesidad de mejorar la eficiencia en el manejo de la información y reducir los costos operativos asociados al consumo de combustible, a través de un modelo que predice la cantidad de combustible consumido en un periodo de tiempo dado.

2) Entendimiento de los datos

Los datos que se tienen para este proyecto son los siguientes:

- **Cantidad de automotores:** La empresa cuenta con 3 vehículos y 1 moto.
- **Recibos de máquina por la compra de combustible:** Estos recibos proporcionan información detallada sobre la compra, incluyendo la cantidad de combustible, el precio por galón, el valor total, la fecha, el nombre del establecimiento y la placa del vehículo.
- **Registro en libros de los recibos:** Este registro proporciona información adicional, incluyendo el nombre del instructor a cargo de la compra y la fecha.
- **Cronograma para asignar las rutas a los instructores:** En este cronograma se identifica al instructor, el vehículo y el tiempo estimado en la ruta para cada estudiante.

3) Preparación de los datos

Se extrajo la información de todos los datos correspondientes a dos meses de este año, octubre y noviembre, donde se segmenta la información recaudada por semanas. Para ello, se

clasifica la información en dos tablas: "Recibos de Maquina" y "Registro en Libros". En la primera se ha clasificado por placa del vehículo y en la segunda por instructor, teniendo en cuenta que a cada uno se le asigna un vehículo en particular, esto con el fin de mantener el orden.

Tablas de Recibos de Máquina: Esta tabla recopila la información de todos los recibos proporcionados por los instructores al comprar combustible para cada carro. Incluye datos de toda una semana por placa, como el número de semana, el rango de semanas correspondiente a los 6 días de la semana (de lunes a sábado), la placa del vehículo, la cantidad total de combustible en galones consumidos y el valor total del gasto semanal.

Tabla de Registro en Libro: En esta tabla se establece el número de semana, el rango de la semana, el nombre del instructor y el valor total gastado en el vehículo.

Tabla 1: Tabla Recibo de Máquina, placa TSY516

TABLA RECIBOS DE MAQUINA				
# Semana	**Rango en días**	**Placa Vehículo**	**Cantidad Combustible Galones**	**Valor Total Compra**
1	02/10 al 07/10	TSY516	5,1	86.700
2	09/10 al 14/10		4	68.000
3	16/10 al 21/10		3,9	66.300
4	23/10 al 28/10		2,3	39.100
5	30/10 al 04/11		3	51.000
6	07/11 al 11/11		4,6	78.200
7	14/11 al 18/11		5,8	98.600
8	20/11 al 25/11		3,5	59.500
9	27/11 al 30/11		1,5	25.500

Tabla 2: Tabla Recibo de Máquina, placa STQ223

TABLA RECIBOS DE MAQUINA				
# Semana	**Rango en días**	**Placa Vehículo**	**Cantidad Combustible Galones**	**Valor Total Compra**
1	02/10 al 07/10	STQ223	2,6	44.200

2	09/10 al14/10		3,5	59.500
3	16/10 al 21/10		3,9	66.300
4	23/10 al28/10		3,6	61.200
5	30/10 al 04/11		3,7	62.900
6	07/11 al 11/11		4	68.000
7	14/11 al 18/11		3,4	57.800
8	20/11 al 25/11		2,2	37.400
9	27/11 al 30/11		2	34.000

Tabla 3: Tabla Recibo de Máquina, placa KDV753

TABLA RECIBOS DE MAQUINA				
# Semana	**Rango en días**	**Placa Vehículo**	**Cantidad Combustible Galones**	**Valor Total Compra**
1	02/10 al 07/10	KDV753	4,5	76.500
2	09/10 al 14/10		4	68.000
3	16/10 al 21/10		4,2	71.400
4	23/10 al 28/10		4,5	76.500
5	30/10 al 04/11		4,8	81.600
6	07/11 al 11/11		4,9	83.300
7	14/11 al 18/11		5	85.000
8	20/11 al 25/11		5,5	93.500
9	27/11 al 30/11		4,4	74.800

Tabla 4: Tabla Recibo de Máquina, placa TIB45G

TABLA RECIBOS DE MAQUINA				
# Semana	**Rango en días**	**Placa Vehículo**	**Cantidad Combustible Galones**	**Valor Total Compra**
1	02/10 al 07/10	TIB45G	1,5	25.500
2	09/10 al 14/10		0,8	13.600
3	16/10 al 21/10		2	34.000

4	23/10 al 28/10		1,7	28.900
5	30/10 al 04/11		0,6	10.200
6	07/11 al 11/11		3	51.000
7	14/11 al 18/11		0,5	8.500
8	20/11 al 25/11		1,2	20.400
9	27/11 al 30/11		0,8	13.600

Tabla 5: Tabla Registro en Libros, Placa TSY516

TABLA REGISTRO EN LIBROS			
# Semana	**Rango en días**	**Nombre Instructor**	**Valor Total de recibo**
1	02/10 al 07/10	**Julián Diaz**	86,700
2	09/10 al 14/10		68,000
3	16/10 al 21/10		66,300
4	23/10 al 28/10		39,100
5	30/10 al 04/11		51,000
6	07/11 al 11/11		78,200
7	14/11 al 18/11		98,600
8	20/11 al 25/11		59,500
9	27/11 al 30/11		25,500

Tabla 6: Tabla Registro en Libros, Placa STQ223

TABLA REGISTRO EN LIBROS			
# Semana	**Rango en días**	**Nombre Instructor**	**Valor Total de recibo**
1	02/10 al 07/10	**Juan Camilo Orduz**	44200
2	09/10 al 14/10		59500
3	16/10 al 21/10		66300
4	23/10 al 28/10		61200

5	30/10 al 04/11		62900
6	07/11 al 11/11		68000
7	14/11 al 18/11		57800
8	20/11 al 25/11		37400
9	27/11 al 30/11		34000

Tabla 7: Tabla Registro en Libros, Placa KDV753

TABLA REGISTRO EN LIBROS			
# Semana	**Rango en días**	**Nombre Instructor**	**Valor Total de recibo**
1	02/10 al 07/10	**Carlos Andrés Leal**	76500
2	09/10 al 14/10		68000
3	16/10 al 21/10		71400
4	23/10 al 28/10		76500
5	30/10 al 04/11		81600
6	07/11 al 11/11		83300
7	14/11 al 18/11		85000
8	20/11 al 25/11		93500
9	27/11 al 30/11		74800

Tabla 8: Tabla Registro en Libros, Placa TIB45G

TABLA REGISTRO EN LIBROS			
# Semana	**Rango en días**	**Nombre Instructor**	**Valor Total de recibo**
1	02/10 al 07/10		25500
2	09/10 al 14/10		13600
3	16/10 al 21/10		34000
4	23/10 al 28/10		28900
5	30/10 al 04/11		10200
6	07/11 al 11/11		51000

7	14/11 al 18/11		8500
8	20/11 al 25/11		20400
9	27/11 al 30/11		13600

Como se puede denotar en las anteriores tablas de datos, encontramos una serie valores del costo total del combustible, la cantidad de galones suministrados para cada vehículo asignado a cada instructor, y su varianza en unas cuantas semanas correspondientes a 2 meses, octubre y noviembre, de esto será necesario definir el modelo para poder determinar las fluctuaciones que se han presentado durante esas semanas, para poder concluir metas claras para el buen manejo de la planificación presupuestaria del consumo de combustible.

4) Generación de código Python para el modelo de clasificación

Se organizan los datos y por lo tanto se generar el modelo predictivo para la estimación del consumo de combustible en una semana, para ello, se utilizaron técnicas de regresión, específicamente regresión lineal múltiple. La variable dependiente o a predecir será el "valor total del recibo", y se consideraron variables independientes como la cantidad de combustible consumido en una semana, la placa del vehículo y el nombre del instructor, el número de semanas y el precio por galón. La relación entre estas variables permite al modelo hacer predicciones sobre el valor total del recibo en función de la cantidad de combustible consumido, el vehículo utilizado y el instructor a cargo. La inclusión del valor total del recibo como variable dependiente es crucial para predecir de manera más precisa los costos asociados al consumo de combustible en la Escuela de Automovilismo Continental del Norte.

Tabla 9. Consolidado de Datos

Valor_Total_Recibo	Cantidad_Combustible_Galones	Nom_Instructor	Placa	Num_Semana	Valor_galon
86700	5.1	Julián Diaz	TSY516	1	17000
44200	2.6	Juan Camilo Orduz	STQ223	3	17000
76500	4.5	Carlos Andrés Leal	KDV753	4	17000
68000	4	Julián Diaz	TSY516	2	17000
59500	3.5	Juan Camilo Orduz	STQ223	2	17000
66300	3.9	Julián Diaz	TSY516	3	17000
71400	4.2	Carlos Andrés Leal	KDV753	3	17000
61200	3.6	Juan Camilo Orduz	STQ223	4	17000
39100	2.3	Julián Diaz	TSY516	4	17000
76500	4.5	Carlos Andrés LEal	KDV753	1	17000
62900	3.7	Juan Camilo Orduz	STQ223	5	17000
78200	4.6	Julián Diaz	TSY516	6	17000
68000	4	Carlos Andrés Leal	KDV753	2	17000
44200	2.6	Juan Camilo Orduz	STQ223	1	17000
81600	4.8	Carlos Andrés LEal	KDV753	5	17000
51000	3	Julián Diaz	TSY520	5	17000
68000	4	Juan Camilo Orduz	STQ223	6	17000
83300	4.9	Carlos Andrés Leal	KDV753	6	17000
98600	5.8	Julián Diaz	TSY523	7	17000
57800	3.4	Juan Camilo Orduz	STQ223	7	17000

5) Aplicación de Códigos para el Modelado

```
# 1 Importar librerias
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from google.colab import drive
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 2 Montar Google Drive
drive.mount ('/content/drive')

# 3 Ruta del archivo de datos en Google Drive
ruta_excel = '/content/drive/MyDrive/Prediccion_Combustible/datos_combustible.xlsx'

# 4 Leer los datos con Pandas
df_Consolidado_datos = pd.read_excel(ruta_excel, sheet_name='Consolidado_datos')

# 5 variables independientes
X = df_Consolidado_datos[['Cantidad_Combustible_Galones', 'Nom_Instructor', 'Placa', 'Num_Semana', 'Valor_galon']]

# 6 variable dependiente
y = df_Consolidado_datos['Valor_Total_Recibo']

# 7 Codificar variables categóricas con one-hot encoding
X_encoded = pd.get_dummies(X, columns=['Nom_Instructor', 'Placa'])

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

# 9 Crear el modelo de regresión lineal múltiple
modelo = LinearRegression()

# 10 Entrenar el modelo con los datos de entrenamiento
modelo.fit(X_train, y_train)

# 11 Realizar predicciones en el conjunto de prueba
predicciones = modelo.predict(X_test)

# 12 Calcular el error cuadrático medio en el conjunto de prueba
error_cuadratico_medio = mean_squared_error(y_test, predicciones)
print(f'Error cuadrático medio: {error_cuadratico_medio}')

# 13 Imprimir los coeficientes y el término independiente del modelo
print('Coeficientes:', modelo.coef_)
print('Término independiente:', modelo.intercept_)
```

Este código realiza un análisis de regresión lineal múltiple para predecir el consumo de combustible para un vehículo, dadas unas variables. El algoritmo comienza importando las

bibliotecas necesarias, como sklearn, pandas, seaborn y matplotlib, que son herramientas esenciales para el análisis y la visualización de datos. Luego, procede a montar Google Drive para acceder a los datos contenidos en el archivo "datos_combustible.xlsx" de la carpeta "MyDrive/Prediccion_Combustible" de Google Drive, estos se leen y se dividen en conjuntos de entrenamiento y de prueba. Esta división facilita el entrenamiento del modelo y la evaluación de su rendimiento, garantizando un análisis exhaustivo. Es importante señalar que las variables categóricas, como el nombre del instructor y la matrícula del vehículo, se codifican utilizando la codificación de un punto, una técnica que transforma estas variables en formatos numéricos comprensibles para el modelo de regresión lineal. Esta la encontramos a partir del comentario 7.

Por otro lado, se crea el modelo de regresión lineal utilizando la clase LinearRegression() de sklearn, lo que demuestra el uso de herramientas especializadas para la construcción de modelos estadísticos. Posteriormente, el modelo se estima con los datos de entrenamiento, y se hacen predicciones sobre el conjunto de pruebas para evaluar su rendimiento.

Por último, se mide la eficacia del modelo, calculando el error cuadrático medio en el conjunto de prueba mediante la función mean_squared_error () de sklearn, proporcionando una métrica cuantitativa sobre la precisión de las predicciones. Los coeficientes y el intercepto del modelo se imprimen en la consola, proporcionando una comprensión detallada de los factores que influyen en las predicciones realizadas, cuyo resultado vemos en la siguiente figura.

```
Error cuadrático medio: 1.0587911840678754e-22
Coeficientes: [ 1.70000000e+04 -5.93071622e-13 -8.07793567e-28  1.36444577e-12
  7.92648402e-13 -4.08366090e-12  1.92656672e-12  1.66673873e-12
  0.00000000e+00  2.59827993e-13  2.15709418e-12 -4.08366090e-12]
Término independiente: 4.3655745685100555e-11
```

Fig. 8. Resultado MSE

Estos resultados son muy positivos y sugieren que el modelo de regresión lineal está funcionando bien en el conjunto de prueba. La baja MSE y los coeficientes pequeños indican que el modelo está capturando las relaciones entre las variables de manera efectiva y está haciendo predicciones precisas.

6) Matriz de Correlación

```
# Matriz de correlación para dos variables

# Seleccionar dos variables específicas
variable1 = 'Cantidad_Combustible_Galones'
variable2 = 'Valor_Total_Recibo'

# Calcular la correlación entre las dos variables
correlación = df_Consolidado_datos[variable1]. corr(df_Consolidado_datos[variable2])
print (f"La correlación entre {variable1} y {variable2} es: {correlación}")
```

En el proceso de desarrollo del modelo predictivo de Machine Learning, se llevó a cabo un análisis detallado de la relación entre las variables clave: Cantidad_Combustible_Galones y Valor_Total_Recibo. Este análisis se realizó mediante la construcción de una matriz correlacional, una herramienta que permite cuantificar la fuerza y dirección de la relación lineal entre dos variables.

La matriz correlacional proporciona coeficientes de correlación que varían entre -1 y 1. Una correlación de 1.0 indica una relación perfecta positiva, lo que significa que a medida que una variable aumenta, la otra también lo hace de manera proporcional. En términos prácticos, una correlación de 1.0 es poco común y sugiere una relación directa y lineal entre las dos variables en el conjunto de datos.

Al examinar la correlación entre Cantidad_Combustible_Galones y Valor_Total_Recibo, se buscó comprender la fuerza y dirección de la relación entre la cantidad de combustible consumido y el valor total del recibo. Un valor alto de correlación positiva indicaría que, en general, a medida que la cantidad de combustible aumenta, el valor total del recibo también tiende a aumentar.

Por otro lado, se determinó que instructor gastó en más combustible en esas semanas transcurridas, mediante una gráfica de dispersión con líneas de barras, que representan el promedio del valor total del recibo para cada instructor de la academia de manejo, lo que fue posible con el siguiente código:

 # Seleccionar solo las columnas relevantes para la gráfica

```
df_Consolidado_datos = pd.read_excel(ruta_excel)
df_grafica = df_Consolidado_datos[['Nom_Instructor', 'Valor_Total_Recibo']]

# Calcular el promedio del valor total del recibo para cada instructor
promedio_valor_por_instructor =
df_grafica.groupby('Nom_Instructor')['Valor_Total_Recibo'].mean().reset_index()

# Crear la gráfica de dispersión
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Nom_Instructor', y='Valor_Total_Recibo', data=df_grafica, alpha=0.7,
label='Datos individuales')
sns.barplot(x='Nom_Instructor',y='Valor_Total_Recibo') data=promedio_valor_por_instructor,
color='purple', label='Promedio por Instructor')
plt.title('Relación entre Instructor y Valor Total del Recibo')
plt.xlabel('Instructor')
plt.ylabel('Valor Total del Recibo')
plt.legend()
plt.show()
```

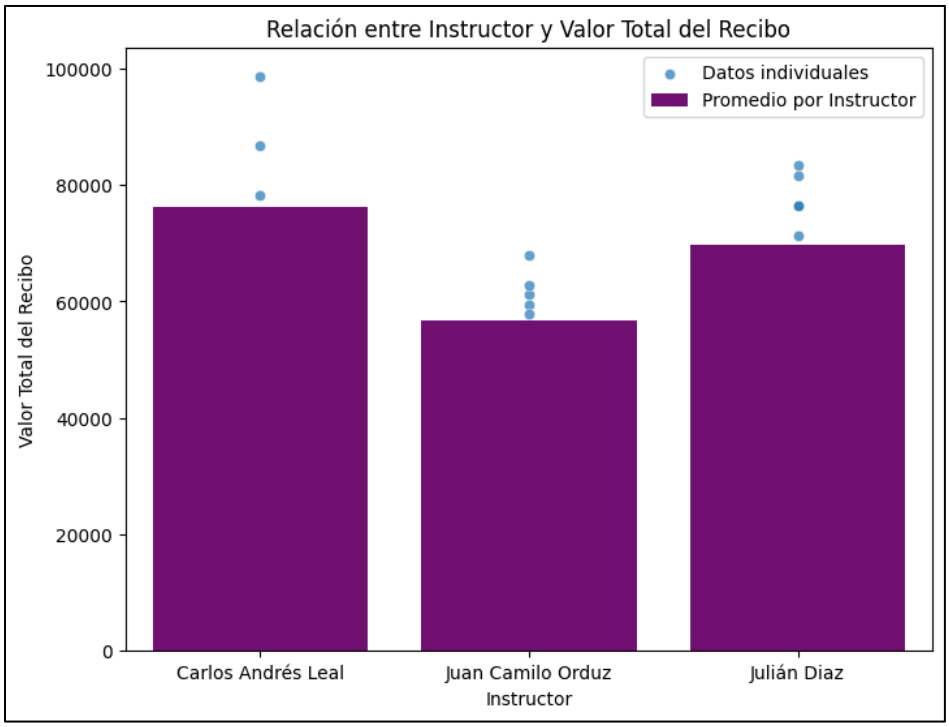


Fig. 9. Gráfica de Dispersión con Líneas de Barras.

Teniendo presente la conexión que hay entre dichas variables además de los otros datos de estudio como el valor del galón de combustible, las semanas transcurridas en las que se suministró combustible al vehículo, y los instructores a los que se le asignó dicho vehículo se puede determinar que el instructor Juan Camilo Orduz, consumió menos combustible durante esas semanas, mientras que los demás tuvieron un gasto más notorio de combustible, lo que implica que la gestión del presupuesto para este tipo de gastos no ha sido la más efectiva en todo su sentido.

De esto se puede derivar las siguientes recomendaciones para una efectiva planificación del presupuesto en combustible:

1. Planificar las clases prácticas de manera eficiente para minimizar los desplazamientos innecesarios y maximizar la utilización de cada vehículo.
2. Llevar un registro detallado para hacer seguimiento preciso del consumo de combustible diario por vehículo, instructor y periodo de tiempo.

3. Agrupar las lecciones en áreas geográficas cercanas, para reducir los kilómetros recorridos.
4. Realización de un mantenimiento preventivo regular en los vehículos para asegurar un rendimiento óptimo y eficiente del combustible.
5. Considerar la posibilidad de incorporar vehículos más eficientes en combustible o incluso opciones eléctricas si es viable desde la parte económica.
6. Buscar acuerdos de exclusividad con proveedores, para obtener tarifas más favorables.

Con lo anterior, se puede determinar que la implementación de estas sugerencias podría y puede contribuir a una gestión más eficiente del presupuesto de combustible, dentro de la Escuela de Automovilismo Continental del Norte, así se estarían optimizando los recursos y promoviendo prácticas sostenibles.

III. CONCLUSIONES

Se identificaron limitaciones cruciales a falta de algunas variables claves, como por ejemplo el precio específico por galón de gasolina, la estación de gasolina a quien se le compraba, la distancia recorrida por semana de cada vehículo, lo cual implica la instalación de GPS en sus automotores, entre otros. Además, no se tuvieron en cuenta las variables inobservables como por ejemplo factores climáticos, puesto que al presentarse días muy soleados o lluviosos, obligan a los conductores a encender el aire acondicionado, ya sea para refrescarse o para evitar que se empañen los vidrios, dicha situación hace que haya más consumo del combustible, todas estas variables no fueron consideradas en el modelo. La ausencia de estas variables adicionales impide una predicción más precisa y completa del consumo de combustible.

La falta de una estructura organizativa y la trazabilidad dentro de la entidad pueden generar dificultades en la hora de salvaguardar los datos y por ende no se pueda generar un modelado de Machine Learning correctamente. Esto se debe a que puede dificultar la identificación de los datos que son relevantes para el modelo, la detección de errores en los datos y la comprensión de la relación entre los datos y la variable objetivo. Para que este modelo predictivo de consumo de combustible sea eficaz, es importante que los datos utilizados para su entrenamiento sean de alta calidad. Esto significa que los datos deben ser precisos, completos y consistentes.

La Escuela de Automovilismo Continental del Norte se enfrenta al desafío de gestionar eficientemente el consumo de combustible de los vehículos que utiliza para dirigir sus clases de conducción. Por lo tanto, la implementación de este modelo permite mejorar la eficiencia en los costos asociados al consumo de combustible de sus vehículos utilizados para las clases de conducción.

La implementación de un modelo de Machine Learning implica costos asociados a la adquisición de tecnologías, formación del personal y gestión de datos. De acuerdo a la estructura que presenta la escuela, en cuanto al orden administrativo, se observa la poca disposición para

suministrar recursos económicos y llevar a cabo este proyecto. Aun teniendo en cuenta que esta inversión puede traducirse en eficiencias a largo plazo y una toma de decisiones más informada.

Se debe llevar una relación respetuosa dentro de la organización, donde incluya la comunicación asertiva y la colaboración entre los instructores y el equipo administrativo, dado que se ha convertido en una parte esencial para lograr resultados óptimos y así la información que se solicite llegue de manera clara y completa.

REFERENCIAS

[1] L. A. Lee, «Machine Learning en tu día a día,» Crehana, Agosto 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11806/player/45886/?source_page=Course%20Dashboard&source_detail=Header%20Button.

[2] L. A. Lee, «La Data a Veces es Sexy,» Crehana, Agosto 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11806/player/45887/.

[3] L. A. Lee, «Entonces...¿Qué es Machine Learning y Por qué Ahora?,» Crehana, Agosto 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11806/player/45888/.

[4] L. A. Lee, «Desmitificando...1,» Crehana, agosto 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11806/player/45890/.

[5] C. Tabares, «¿Qué es un contrafactual?,» Crehana, Septiembre 2022. [En línea]. Available: https://whttps://www.crehana.com/clases/v2/16430/player/64337/?source_page=Course%20Dashboard&source_detail=Header%20Button.

[6] C. Tabares, «Correlación vs. Causalidad,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64338/.

[7] C. Tabares, «Variables Omitidas y Sesgo de Selección,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64339/.

[8] C. Tabares, «Propensity Score,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64350/.

[9] C. Tabares, «Double LASSO,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64351/.

[10] C. Tabares, «Causal Trees,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64353/.

[11] C. Tabares, «Casual Forest,» Crehana, Septiembre 2022. [En línea]. Available: https://www.crehana.com/clases/v2/16430/player/64354/.

[12] M. Rojo, «Machine Learning Workflow,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69013/?source_page=Course%20Dashboard&source_detail=Header%20Button.

[13] M. Rojo, «Preparación de Datos parte 1,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69014/.

[14] M. Rojo, «Preparación de Datos parte 2,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69015/.

[15] M. Rojo, «Modelado,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69016/.

[16] M. Rojo, «Matriz de Confusión,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69019/.

[17] M. Rojo, «Accuracy,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69020/.

[18] M. Rojo, «Precisión y Recall,» Crehana , Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69021/.

[19] M. Rojo, «Specifcity y F1 Score,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69022/.

[20] M. Rojo, «Curva de Precision-Recall,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69024/.

[21] M. Rojo, «Técnicas de Validación Cruzada,» Crehana, Enero 2023. [En línea]. Available: https://www.crehana.com/clases/v2/16963/player/69026/.

[22] H. A. Aragón, «Inteligencia Artificial, Machine Learning Supervisado y No Supervisado,» Crehana, Octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/12043/player/48349/.

[23] H. A. Aragón, «Regresión Lineal,» Crehana, Octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/12043/player/48357/.

[24] H. A. Aragón, «Regresión Logística,» Crehana, Octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/12043/player/48358/.

[25] H. A. Aragón, «Árboles de Decisión - Continuos,» Crehana, Octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/12043/player/48359/.

[26] H. A. Aragón, «Introducción Neural Network,» Crehana, Octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/12043/player/48366/.

[27] L. A. Lee, «Python, paquetes y librerias,» Crehana, 15 octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11939/player/47928/.

[28] L. A. Lee, «Analiza las variables 2,» Crehana, 15 octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11939/player/47931/?t=332.

[29] L. A. Lee, «¿Con o sin supervision?,» Crehana, 15 octubre 2021. [En línea]. Available: https://www.crehana.com/clases/v2/11939/player/47943/?t=49.

[30] F. D. Ruiz Martinez, «Estado del Arte de la Inteligencia Artificial,» Crehana, enero 2022. [En línea]. Available: https://www.crehana.com/clases/v2/12531/player/51143/.

[31] F. D. Ruiz Martinez, «Diferencias entre AI, ML, DL y Data Science,» Crehana, Enero 2022. [En línea]. Available: https://www.crehana.com/clases/v2/12531/player/51145/?source_page=Course%20Dashboard&source_detail=Content%20section.

[32] F. D. Ruiz Martinez, «Arquitectura de las redes Neuronales,» Crehana, enero 2022. [En línea]. Available: https://www.crehana.com/clases/v2/12531/player/51149/.

[33] E. Wohlmuth, «Conceptos fundamentales de ética aplicados al campo de la inteligencia artificial.,» Crehana, Agosto 2022. [En línea]. Available: https://www.crehana.com/clases/v2/25688/player/80387/.

[34] R. A. Murga Garrido, «Beneficios de la IA en la innovacion tecnologica,» Crehana, septiembre 2023. [En línea]. Available: https://www.crehana.com/clases/v2/25736/player/81474/.

[35] R. A. Murga Garrido, «Diseño de soluciones innovadoras y tecnologicas con IA - Parte II,» Crehana, septiembre 2023. [En línea]. Available: https://www.crehana.com/clases/v2/25736/player/81477/.

APENDICE FIGURAS Y TABLAS