



Aplicación de soluciones a negocios

Equipo 22

Daniela Gonzalez Cruz
José De Jesús Barajas Figueroa
Edgardo Gonzalez Zepeda
Luis Manuel Espinoza Juárez
José Eduardo González Moreno



BEDU - Santander

Experimentando con
herramientas sin código

Antecedentes

Nuestra empresa, **AICommerce**, es una consultoría especializada en ayudar a empresas extranjeras a establecerse en México, utilizando Inteligencia Artificial para solución de desafíos relacionados con ventas. Actualmente, nos encontramos brindando servicio a la empresa *Foreign Cars* para facilitar su entrada en el muy competitivo mercado automotriz mexicano. Una de sus principales inquietudes es conocer el precio óptimo al que venderán sus modelos, dado que no tienen conocimiento del precio que manejan sus futuros competidores ni los criterios en los que se basan para establecerlo.

Por ello, **AICommerce** está desarrollando un modelo de Machine Learning capaz de analizar los precios de los automóviles en función de características como dimensiones, especificaciones del motor y calificaciones de seguridad. Este modelo proporcionará a *Foreign Cars* una estimación de los precios de sus vehículos, basada en los datos de los competidores ya establecidos en el país.

Background de la Empresa

AICommerce es una consultoría innovadora especializada en el uso de tecnologías avanzadas de Inteligencia Artificial (IA) para ayudar a empresas con un background no especializado y de perfil no técnico, con soluciones hechas a la medida.



AICommerce fue fundada por un grupo de expertos en IA y análisis de datos que en conjunto suman décadas de experiencia en la industria. Esto le ha permitido a **AICommerce** posicionarse rápidamente como líder en la integración de soluciones tecnológicas en el ámbito empresarial. La empresa se dedica a optimizar los procesos de entrada al mercado, ofreciendo a sus clientes una ventaja competitiva en sectores altamente dinámicos y competitivos, como el automotriz. Gracias a su enfoque centrado en la IA, **AICommerce** es capaz de transformar grandes volúmenes de datos en insights valiosos que facilitan la toma de decisiones estratégicas.

Descripción del Problema

Una empresa de distribución automovilista de renombre y consolidada en países con economías más estables y desarrollos tecnológicos automotrices de vanguardia desea entrar al mercado mexicano, para esto se necesita evaluar su oferta de productos, capacidades, mercado objetivo, y los elementos que ofrecen sus vehículos en comparación de los que ya existen, tanto en producción como en importación en el mercado mexicano actual y con esto poder crear un sistema de precios adecuado y competitivo.

En este caso específico abordaremos el problema con *Foreign Cars*, una compañía automotriz internacional. La determinación del precio adecuado para sus modelos es crucial, ya que influye directamente en su competitividad y en la percepción del valor de sus productos por parte de los consumidores mexicanos. Por ello es que, sin una comprensión clara de los factores que determinan los precios de los competidores,

Foreign Cars está en riesgo de salir al mercado con precios que podrían ser inadecuados, y dado que sería la primer impresión financiera de su producto no será posible hacer una estrategia de revalorización sin perder clientes, o perder aceptación comercial pública.

Para abordar este problema, **AlCommerce** está desarrollando un modelo de Machine Learning que analiza una serie de variables, incluyendo las dimensiones de los vehículos, las especificaciones del motor, calificaciones de seguridad, entre otros elementos mecánicos y de confort al usuario, para sugerir precios óptimos en sus productos. Este modelo permitirá a *Foreign Cars* tomar decisiones informadas sobre su estrategia de precios, basándose en datos concretos del mercado automotriz mexicano. Al contar con esta herramienta, *Foreign Cars* podrá tener una entrada más efectiva y rentable en México.

[Como ya sabemos: A diferencia de la programación tradicional, donde se definen reglas explícitas para el procesamiento de datos, los modelos de ML utilizan un paradigma de aprendizaje. En este paradigma, los modelos de ML se entrenan con un conjunto de datos conocido como "conjunto de entrenamiento". El modelo aprende a partir de estos ejemplos y utiliza este conocimiento para realizar predicciones o tomar decisiones en nuevos casos.]

Minería de datos

Con el fin de generar un modelo que atienda las necesidades de la empresa se decidió utilizar un [dataset](#) de corte transversal obtenido de Kaggle, el cual recopila información de autos con su precio de venta, los precios considerados en la tabla tienen un corte al año 2023.

El dataset cuenta con columnas, las cuales son explicadas a continuación :

- **CarID:** Número de identificación de cada coche
- **SR:** (Safety Rating) Calificación de seguridad del coche
- **FT:** (Fuel Type) Tipo de combustible utilizado (gasolina, diesel, eléctrico, etc.)
- **Asp:** (Aspiration) Tipo de aspiración (estándar o turbo)
- **ND:** (Number of Doors) Número de puertas del coche
- **BS:** (Body Style) Estilo de la carrocería del coche (Sedán, Coupé, SUV, etc.)
- **DWT:** (Drive Wheel Type) Tipo de ruedas motrices (Delanteras, Traseras, Todas)
- **EnL:** (Engine Location) Ubicación del motor del coche (Delantero o Trasero)
- **CL:** (Car Length) Longitud total del vehículo
- **CW:** (Car Width) anchura del vehículo
- **CH:** (Car Height) altura del vehículo
- **WCW:** (Curb Weight) Peso del vehículo sin pasajeros ni carga
- **EnT:** (EngineType) Tipo de motor Tipo de motor (gasolina, diesel, eléctrico, etc.)
- **NC:** (Number Cylinders) Número de cilindros del motor
- **EnS:** (Engine Size) Tamaño del motor del coche
- **FSys:** (Fuel System) Tipo de sistema de suministro de combustible
- **HP:** (Horse Power) Potencia del motor del coche
- **P-RPM:** (Peak RPM) RPM máximas del motor (revoluciones por minuto)
- **C-KPL:** (City KPL) Kilómetros por Litro en Ciudad
- **H-KPL:** (Highway KPL) Kilómetros por Litro en Pista
- **CP:** (Car Price) Precio del coche

Esta codificación se realizó adaptando el dataset original de Knime para tener una nomenclatura interna, así mismo poder abreviar en el texto cada uno de los atributos.

Identificación de Soluciones IA y ML

Los modelos de Machine Learning e Inteligencia Artificial son herramientas poderosas capaces de beneficiar y optimizar la toma de decisiones en los negocios. Esto, ya que permiten identificar patrones y tendencias a partir de grandes cantidades de datos, facilitando la creación de estrategias para la venta de productos y ofreciendo el precio óptimo por ellos.

Para ayudar a Foreign Cars a establecerse exitosamente en el mercado automotriz mexicano y poder fijar el precio óptimo para sus modelos, **AlCommerce** propone varias soluciones de Inteligencia Artificial (IA) y Machine Learning (ML). Aquí hay algunas opciones que pueden beneficiar a la empresa:

1. **Modelo de Regresión lineal para Predicción de Precios:** Creación de un modelo de regresión lineal para predecir el precio de los automóviles basándose en las características y los precios de los competidores en el mercado, para que puedan competir exitosamente.
2. **Análisis de Datos de la Competencia con Procesamiento de Lenguaje Natural (NLP):** Implementación de técnicas de NLP para analizar y extraer información de los sitios web y fichas técnicas de los vehículos de la competencia, transformando de datos no estructurados (precios y características) a una base de datos estructurada.
3. **Segmentación de Mercado con Clustering:** Uso de técnicas de clustering (como K-means) para identificar grupos según características demográficas y preferencias, y así poder personalizar estrategias de marketing para cada segmento del mercado.

Implementando estas tres soluciones, se lograría una mayor precisión en la fijación de los precios de los automóviles. Con la primera solución se desarrollaría un modelo de IA escalable, que se alimentará de la información recopilada constantemente por la segunda solución, el Análisis de Datos de la Competencia con NLP. Esto permitiría que los datos permanezcan actualizados y lograría una comparación continua con la competencia. Además, al definir las preferencias de cada segmento del mercado con la tercera solución, se podrían ajustar los precios según la cantidad que cada grupo esté dispuesto a pagar.

Estas soluciones beneficiarán a la empresa minimizando el riesgo de establecer precios demasiado altos o bajos, maximizando su competitividad, rentabilidad y percepción por parte del público, posicionando así a la marca en el mercado mexicano.

Ejecución de la Solución

Para fines prácticos de este proyecto, solo se implementará la primera solución propuesta: el **Modelo de Regresión para Predicción de Precios**. Esta elección se debe a que se adapta de mejor manera a las necesidades inmediatas de Foreign Cars y ofrece un enfoque claro y efectivo para la entrada al mercado.

Desarrollo de Algoritmo

¿Por qué regresión lineal?

La elección de un modelo de regresión se basa en su capacidad para identificar relaciones directas entre las variables independientes (como las características de los vehículos) y la variable dependiente (el precio). Por ello, AICommerce considera que este modelo es ideal para este proyecto debido a su simplicidad y efectividad en la predicción de precios cuando las relaciones son lineales. Además, los modelos de regresión son altamente interpretativos, lo que facilita la explicación de los resultados a los stakeholders. Recordemos que los stakeholders no tienen un perfil técnico, por lo que en términos de **explicabilidad** es una elección lógica.

El modelo de regresión lineal se construirá utilizando un conjunto de datos históricos de precios y características de vehículos disponibles en el mercado mexicano. Este modelo se entrenará para predecir los precios de los automóviles de Foreign Cars (CP: Car Price), considerando factores como las especificaciones técnicas, las dimensiones y las calificaciones de seguridad. Considerando que los ejecutivos de Foreign Cars no cuentan con un perfil técnico, se optó por implementar el modelo usando KNIME, una herramienta de código abierto que permite realizar modelos de IA sin código por medio de una estructura de nodos.

El modelo de regresión lineal múltiple desarrollado en KNIME tiene el mismo objetivo de predecir el precio de los automóviles, utilizando un conjunto de variables técnicas y características de los vehículos. A continuación, se analizarán los resultados obtenidos del modelo, evaluando su rendimiento a través de las métricas y evaluaciones proporcionadas por KNIME, y explorando la importancia de las variables predictoras en la determinación del precio de los autos.

1. Preprocesamiento de Datos:

Antes de realizar un análisis o un modelado específico, es necesario realizar una labor de limpieza de los datos, estandarización y normalización de las variables que integran los datos. A continuación mencionaremos brevemente las partes más relevantes de este proceso.

- **Codificación de Variables Categóricas:** Variables como **CN** (Car Name), **FT** (Fuel Type), **Asp** (Aspiration), **BS** (Body Style), **DWT** (Drive Wheel Type), **EnL** (Engine Location), **EnT** (Engine Type), y **FSys** (Fuel System) son categóricas y necesitan ser codificadas numéricamente (utilizando los nodos adecuados en KNIME), esto con el objetivo de que puedan ser utilizadas correctamente en el modelo de regresión.
- **Normalización/Estandarización:** Dado que las variables como **CL** (Car Length), **CW** (Car Width), **CH** (Car Height), **WCW** (Curb Weight), **EnS** (Engine Size), **HP** (Horse Power), **P-RPM** (Peak RPM), **C-KPL** (City KPL) y **H-KPL** (Highway KPL) tienen diferentes escalas, es recomendable normalizarlas o estandarizarlas para que tengan un impacto equilibrado en el modelo.
- **Manejo de Valores Faltantes:** Finalmente, es necesario identificar y manejar cualquier valor faltante en el dataset. Esto podría implicar la imputación de valores faltantes o la eliminación de filas/columnas con datos incompletos.

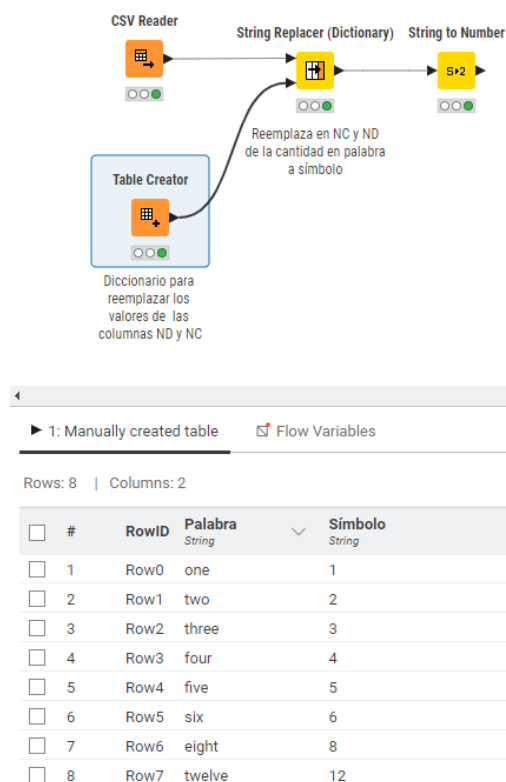
Aterrizando este proceso en Knime:

1. Se utilizó el nodo *CSV Reader* para poder leer la base de datos y poder encontrar algunas peculiaridades en ellos a simple vista. Lo primero de que nos percatamos fue que las variables **NC** (Number of Cylinders) y **ND** (Number of Doors) contenían cantidades numéricas pero con palabras en vez de símbolos, por ejemplo, en vez de contener el número 4 tenía el texto “four”.

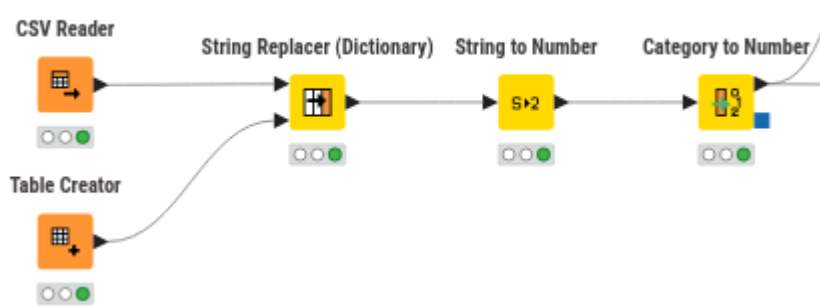
Rows: 205 | Columns: 22

#	RowID	CarID	SR	CN	FT	Asp	ND	BS	DWT	EnL	CL	CW	CH	WCW	EnT	NC
		Numb...	Numb...	String	String	String	String	String	String	String	Number (dou...	Number (dou...	Number (dou...	Number (inte...	String	String
1	Row0	1	3	alfa-romero gi...	gas	std	two	convertible	rwd	front	168.8	64.1	48.8	2548	dohc	four
2	Row1	2	3	alfa-romero st...	gas	std	two	convertible	rwd	front	168.8	64.1	48.8	2548	dohc	four
3	Row2	3	1	alfa-romero Q...	gas	std	two	hatchback	rwd	front	171.2	65.5	52.4	2823	ohcv	six
4	Row3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	176.6	66.2	54.3	2337	ohc	four
5	Row4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	176.6	66.4	54.3	2824	ohc	five
6	Row5	6	2	audi fox	gas	std	two	sedan	fwd	front	177.3	66.3	53.1	2507	ohc	five
7	Row6	7	1	audi 100ls	gas	std	four	sedan	fwd	front	192.7	71.4	55.7	2844	ohc	five
8	Row7	8	1	audi 5000	gas	std	four	wagon	fwd	front	192.7	71.4	55.7	2954	ohc	five
9	Row8	9	1	audi 4000	gas	turbo	four	sedan	fwd	front	192.7	71.4	55.9	3086	ohc	five
10	Row9	10	0	audi 5000s (d...	gas	turbo	two	hatchback	4wd	front	178.2	67.9	52	3053	ohc	five

Para corregir esto y poder seguir con los pasos de preprocesamiento mencionados anteriormente, se creó un diccionario utilizando el nodo *Table Creator* y se reemplazaron los valores con el nodo *String Replacer (Dictionary)*, para después convertir los valores de texto a enteros utilizando el nodo *String to Number*.



2. Se utilizó el nodo *Category to Number* para que las verbales **CN**, **FT**, **Asp**, **BS**, **DWT**, **EnL**, **EnT**, **FSys**, ya que el algoritmo de regresión lineal requiere datos numéricos como entrada



Rows: 205 | Columns: 22

Table Statistics



☐	#	RowID	CarID Number (inte...	SR Number (inte...	CN Number (inte...	FT Number (inte...	Asp Number (inte...	ND Number (inte...	BS Number (inte...	DWT Number (ini
☐	1	Row0	1	3	1	1	1	2	1	1
☐	2	Row1	2	3	2	1	1	2	1	1
☐	3	Row2	3	1	3	1	1	2	2	1
☐	4	Row3	4	2	4	1	1	4	3	2
☐	5	Row4	5	2	5	1	1	4	3	3
☐	6	Row5	6	2	6	1	1	2	3	2
☐	7	Row6	7	1	5	1	1	4	3	2
☐	8	Row7	8	1	7	1	1	4	4	2
☐	9	Row8	9	1	8	1	2	4	3	2
☐	10	Row9	10	0	9	1	2	2	2	3
☐	11	Row10	11	2	10	1	1	2	3	1
☐	12	Row11	12	0	10	1	1	4	3	1
☐	13	Row12	13	0	11	1	1	2	3	1

- Después, se utilizó el nodo *Statistics View* para visualizar las estadísticas básicas de los datos. Gracias a ello, pudimos identificar que no existían valores faltantes en ninguna columna.

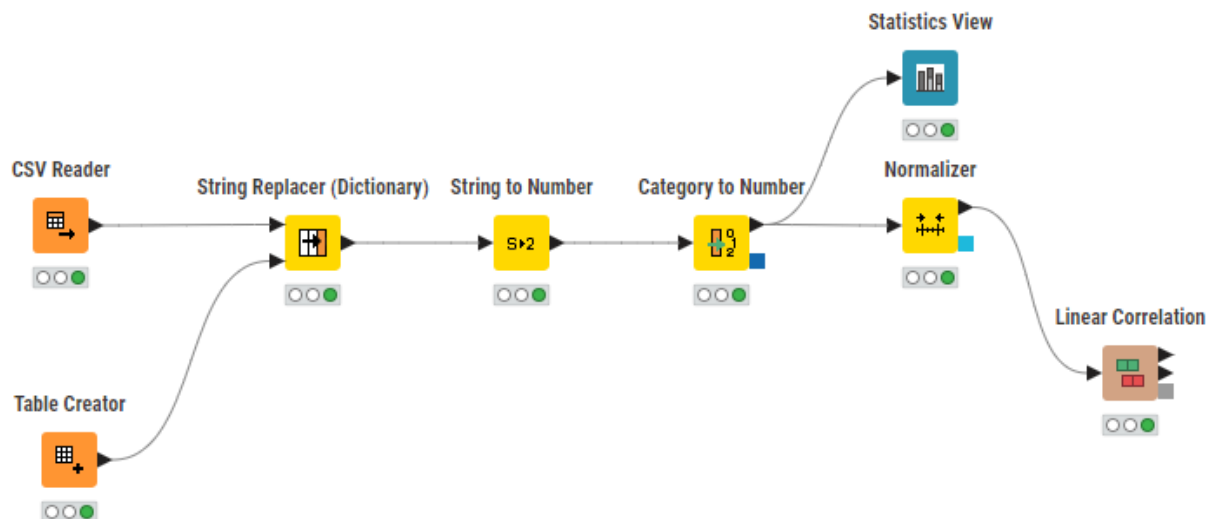
Statistics

Rows: 22 | Columns: 14

Name	Type	# Missing values	# Unique values	Minimum	Maximum	25% Quantile	50% Quantile (Med...)	75% Quantile	Mean	Mean Absolute De...	Standard Deviation	Sum	10 most common ...
CarID	Number (integer)	0	205	1	205	51.5	103	154.5	103	51.249	59.323	21,115	1 (1; 0.49%), 2 (1; 0...
SR	Number (integer)	0	6	-2	3	0	1	2	0.834	1.022	1.245	171	0 (67; 32.68%), 1 (5...
CN	Number (integer)	0	147	1	147	43.5	80	114	77.449	35.633	41.346	15,877	85 (6; 2.93%), 114 (...)
FT	Number (integer)	0	2	1	2	1	1	1	1.098	0.176	0.297	225	1 (185; 90.24%), 2 (...)
Asp	Number (integer)	0	2	1	2	1	1	1	1.18	0.296	0.386	242	1 (168; 81.95%), 2 (...)
ND	Number (integer)	0	2	2	4	2	4	4	3.122	0.985	0.995	640	4 (115; 56.1%), 2 (9...
BS	Number (integer)	0	5	1	5	2	3	3	2.8	0.652	0.837	574	3 (96; 46.83%), 2 (7...
DWT	Number (integer)	0	3	1	3	1	2	2	1.673	0.499	0.556	343	2 (120; 58.54%), 1 (...)
EnL	Number (integer)	0	2	1	2	1	1	1	1.015	0.029	0.12	208	1 (202; 98.54%), 2 (...)
CL	Number (double)	0	75	141.1	208.1	166.3	173.2	183.3	174.049	9.594	12.337	35,680.1	157.3 (15; 7.32%), ...
CW	Number (double)	0	44	60.3	72.3	64.05	65.5	66.9	65.908	1.678	2.145	13,511.1	63.8 (24; 11.71%), ...
CH	Number (double)	0	49	47.8	59.8	52	54.1	55.5	53.725	2.015	2.444	11,013.6	50.8 (14; 6.83%), 5...
WCW	Number (integer)	0	171	1,488	4,066	2,145	2,414	2,943.5	2,555.566	428.268	520.68	523,891	2,385 (4; 1.95%), 1...
EnT	Number (integer)	0	7	1	7	3	3	3	3.156	0.624	1.073	647	3 (148; 72.2%), 6 (1...
NC	Number (integer)	0	7	2	12	4	4	4	4.38	0.697	1.081	898	4 (159; 77.56%), 6 (...)
EnS	Number (integer)	0	44	61	326	97	120	143	126.907	30.224	41.643	26,016	92 (15; 7.32%), 122...
FSys	Number (integer)	0	8	1	8	1	2	2	2.478	1.663	2.184	508	1 (94; 45.85%), 2 (6...
HP	Number (integer)	0	59	48	288	70	95	116	104.117	30.317	39.544	21,344	68 (19; 9.27%), 70 (...)
P-RPM	Number (integer)	0	23	4,150	6,600	4,800	5,200	5,500	5,125.122	382.315	476.986	1,050,650	5,500 (37; 18.05%), ...
C-KPL	Number (integer)	0	29	13	49	19	24	30	25.22	5.215	6.542	5,170	31 (28; 13.66%), 19...
HKPL	Number (integer)	0	30	16	54	25	30	35	30.751	5.467	6.886	6,304	25 (19; 9.27%), 24 (...)
CP	Number (double)	0	189	5,118	45,400	7,781.5	10,295	16,509	13,276.711	5,898.504	7,988.852	2,721,725.667	5,572 (2; 0.98%), 6...

Por lo que no fue necesario implementar nodos que reemplazan valores faltantes o bien eliminar elemento con datos incompletos

- Después, se normalizaron las variables con el nodo *Normalizer* en un rango de 0 a 1, y se visualizó la correlación de los datos generando una matriz de correlación a partir del nodo *Lineal Correlation*.



Al analizar la matriz de correlación se puede identificar que las variables que tienen mayor correlación con el precio (CP) son CL (Car Length), CW (Car Width), WCW (Curb Weight), NC (Number of Cylinders), HP (Horse Power) y EnS (Engine Size), donde esta última tiene el valor más alto.

Row ID	SR	CN	FT	Asp	ND	BS	DWT	EnL	CL	CW	CH	WCW	EnT	NC	EnS	FSys	HP	P-RPM	C-KPL	H-KPL	CP
SR	1	-0.137	-0.194	-0.06	-0.664	-0.399	0.042	0.212	-0.358	-0.233	-0.541	-0.228	0.016	-0.113	-0.106	0.017	0.071	0.274	-0.036	0.035	-0.08
CN	-0.137	1	0.128	0.075	0.178	0.173	-0.045	0.061	0.161	0.044	0.253	0.068	0.118	-0.102	-0.038	-0.048	-0.019	-0.198	0.018	0.014	-0.114
FT	-0.194	0.128	1	0.401	0.191	0.158	-0.132	-0.04	0.213	0.234	0.285	0.217	0.029	-0.025	0.07	0.682	-0.164	-0.477	0.256	0.191	0.106
Asp	-0.06	0.075	0.401	1	0.032	0.021	-0.066	-0.057	0.235	0.301	0.087	0.325	0.015	-0.048	0.108	0.473	0.242	-0.183	-0.202	-0.254	0.178
ND	-0.664	0.178	0.191	0.032	1	0.448	0.099	-0.138	0.399	0.207	0.552	0.197	0.037	-0.016	0.021	-0.018	-0.127	-0.248	-0.012	-0.036	0.032
BS	-0.399	0.173	0.158	0.021	0.448	1	-0.057	0.127	0.377	0.192	0.484	0.24	0.128	0.106	0.172	-0.055	0.047	-0.147	-0.096	-0.127	0.161
DWT	0.042	-0.045	-0.132	-0.066	0.099	-0.057	1	-0.148	-0.486	-0.471	0.02	-0.575	0.193	-0.314	-0.524	0.036	-0.519	0.039	0.45	0.452	-0.578
EnL	0.212	0.061	-0.04	-0.057	-0.138	0.127	-0.148	1	-0.051	-0.052	-0.106	0.05	0.324	0.183	0.197	-0.083	0.318	0.198	-0.153	-0.102	0.325
CL	-0.358	0.161	0.213	0.235	0.399	0.377	-0.486	-0.051	1	0.841	0.491	0.878	-0.172	0.431	0.683	-0.091	0.553	-0.287	-0.671	-0.705	0.683
CW	-0.233	0.044	0.234	0.301	0.207	0.192	-0.471	-0.052	0.841	1	0.279	0.867	-0.098	0.545	0.735	-0.017	0.641	-0.22	-0.643	-0.677	0.759
CH	-0.541	0.253	0.285	0.087	0.552	0.484	0.02	-0.106	0.491	0.279	1	0.296	0.01	-0.014	0.067	-0.051	-0.109	-0.32	-0.049	-0.107	0.119
WCW	-0.228	0.068	0.217	0.325	0.197	0.24	-0.575	0.05	0.878	0.867	0.296	1	-0.165	0.61	0.851	-0.026	0.751	-0.266	-0.757	-0.797	0.835
EnT	0.016	0.118	0.029	0.015	0.037	0.128	0.193	0.324	-0.172	-0.098	0.01	-0.165	1	-0.25	-0.231	0.079	-0.104	-0.055	0.064	0.078	-0.093
NC	-0.113	-0.102	-0.025	-0.048	-0.016	0.106	-0.314	0.183	0.431	0.545	-0.014	0.61	-0.25	1	0.846	-0.229	0.692	-0.124	-0.446	-0.467	0.718
EnS	-0.106	-0.038	0.07	0.108	0.021	0.172	-0.524	0.197	0.683	0.735	0.067	0.851	-0.231	0.846	1	-0.132	0.81	-0.245	-0.654	-0.677	0.874
FSys	0.017	-0.048	0.682	0.473	-0.018	-0.055	0.036	-0.083	-0.091	-0.017	-0.051	-0.026	0.079	-0.229	-0.132	1	-0.226	-0.206	0.262	0.217	-0.122
HP	0.071	-0.019	-0.164	0.242	-0.127	0.047	-0.519	0.318	0.553	0.641	-0.109	0.751	-0.104	0.692	0.81	-0.226	1	0.131	-0.801	-0.771	0.808
P-RPM	0.274	-0.198	-0.477	-0.183	-0.248	-0.147	0.039	0.198	-0.287	-0.22	-0.32	-0.266	-0.055	-0.124	-0.245	-0.206	0.131	1	-0.114	-0.054	-0.085
C-KPL	-0.036	0.018	0.256	-0.202	-0.012	-0.096	0.45	-0.153	-0.671	-0.643	-0.049	-0.757	0.064	-0.446	-0.654	0.262	-0.801	-0.114	1	0.971	-0.686
H-KPL	0.035	0.014	0.191	-0.254	-0.036	-0.127	0.452	-0.102	-0.705	-0.677	-0.107	-0.797	0.078	-0.467	-0.677	0.217	-0.771	-0.054	0.971	1	-0.698
CP	-0.08	-0.114	0.106	0.178	0.032	0.161	-0.578	0.325	0.683	0.759	0.119	0.835	-0.093	0.718	0.874	-0.122	0.808	-0.085	-0.686	-0.698	1

- Se validó mediante el nodo Row Filter que no existieran filas duplicadas, debido a que no se encontró ninguna el flujo no necesariamente requiere seguir el nodo de filtrado utilizado.



Antes de comenzar el entrenamiento se enlistan algunos datos extraídos del análisis del Dataset:

- Para este modelo se está analizando el precio de 205 modelos de auto con características diferentes entre si (no hay ningún carro igual, aun cuando sean de la misma marca)
- Los rangos de los vehículos analizado van de los \$5118.0 a los \$45400.0 dólares, lo cual genera un amplio margen de distintas gamas de precio lo cual es conveniente para nuestro cliente ya que quiere incursionar dentro de varios segmentos en el mercado, siendo el precio promedio de los vehículos en el dataset \$13276.71 dólares..

2. División del Dataset:

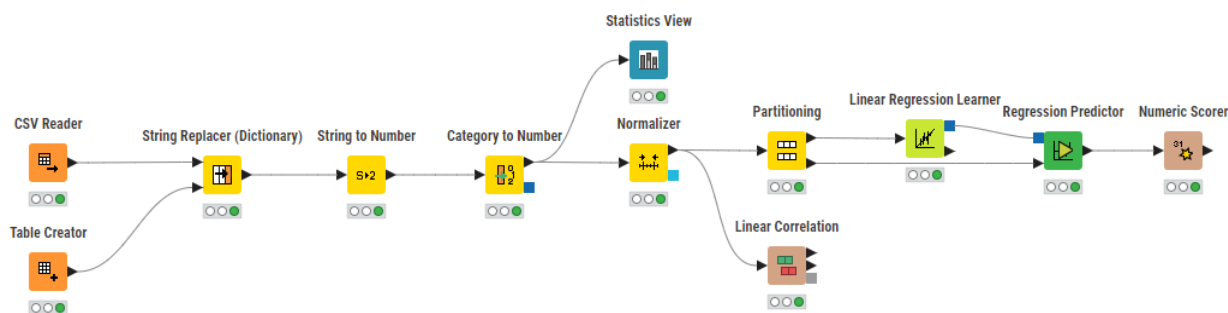
- **Entrenamiento y Prueba:** Inicialmente se realizó la división de los datos utilizando el nodo *Partitioning* en un conjunto de entrenamiento y un conjunto de prueba. Usualmente, se utiliza un 70-80% de los datos para el entrenamiento y el 20-30% restante para la prueba. En este caso utilizamos 70% para entrenamiento y 30% para prueba, ya que fue la combinación que mejores resultados arrojó durante las pruebas. Finalmente, esto permitirá evaluar la capacidad del modelo para generalizar sobre datos no vistos.

3. Construcción del Modelo:

- **Modelo de Regresión Lineal Múltiple:** Inicialmente se emplearon todas las variables predictoras (SR, FT, Asp, ND, CN, BS, DWT, EnL, CL, CW, CH, WCW, EnT, NC, EnS, FSys, HP, P-RPM, C-KPL, H-KPL) para construir un modelo de regresión lineal múltiple que prediga el precio del automóvil (CP), utilizando los nodos *Linear Regression Learner* y *Regression Predictor*.

4. Evaluación del Modelo:

- **Métricas de Evaluación:** Se evaluó el rendimiento del modelo usando el conjunto de datos de prueba y el nodo *Numeric Scorer*. Algunas métricas clave incluyen el R-cuadrado (R^2), el error cuadrático medio (MSE) y el error absoluto medio (MAE).



Para el modelo que contemplaba todas las variables, se obtuvieron los siguientes valores:

RowID	Prediction (CP) <i>Number (double)</i>
R^2	0.864
mean absolute error	0.056
mean squared error	0.005
root mean squared error	0.068
mean signed difference	0.024
mean absolute percentag...	0.507
adjusted R^2	0.864

En una segunda iteración se decidió sacar el valor de **CN** (Car Name) del conjunto de entrenamiento esto porque los autos que *ForeignCars* no comparten nombre con ninguno de los modelos actuales en el mercado y queremos medir el costo del carro por sus características, el nombre de los modelos queda a cargo del área de marketing.

The screenshot shows a variable selection interface. At the top, there's a dropdown menu set to 'D CP'. Below it, the 'Values' section has two tabs: 'Manual Selection' (selected) and 'Wildcard/Regex Selection'. Under 'Manual Selection', there are two main panels: 'Exclude' (outlined in red) and 'Include' (outlined in green). The 'Exclude' panel has a 'Filter' input field and a list containing 'CarID' and 'CN', with 'CN' selected. Below this list is a radio button labeled 'Enforce exclusion' which is selected. The 'Include' panel also has a 'Filter' input field and a list of car model abbreviations: 'SR', 'FT', 'Asp', 'ND', 'BS', 'DWT', 'EnL', 'CL', 'CW', and 'CH'. Below this list is a radio button labeled 'Enforce inclusion' which is not selected. Between the two panels are four navigation buttons: '>', '>>', '<', and '<<'. At the bottom of the interface, the text 'Regression Properties' is visible.

Con este ajuste se vio un cambio significativo en los resultados obtenidos, sobretodo en el factor de R^2 , el cual paso de .864 a .898

☐	#	RowID	Prediction (CP) <i>Number (double)</i>
☐	1	R^2	0.898
☐	2	mean absolute error	0.054
☐	3	mean squared error	0.005
☐	4	root mean squared error	0.069
☐	5	mean signed difference	0.021
☐	6	mean absolute percentage error	NaN
☐	7	adjusted R^2	0.898

5. Interpretación del Modelo:

- **Coefficientes del Modelo:** Analiza los coeficientes de las variables para entender la influencia de cada factor en el precio del coche. Un coeficiente positivo indicará que un aumento en esa variable está asociado con un aumento en el precio, mientras que un coeficiente negativo indicará lo contrario.
- **Importancia de Variables:** Identifica las variables más influyentes en la predicción del precio, lo que podría proporcionar insights adicionales sobre qué características de los coches son más valoradas en el mercado mexicano.

1. R² (R cuadrado): 0.898

El valor de R² indica que aproximadamente el 89.8% de la variabilidad en el precio del coche (CP) puede ser explicada por las variables predictoras incluidas en el modelo. Este es un buen nivel de ajuste, lo que sugiere que el modelo tiene un buen poder predictivo para las observaciones dentro del conjunto de datos utilizado.

2. Mean Absolute Error (MAE): 0.054

El MAE mide la magnitud promedio de los errores en un conjunto de predicciones, sin tener en cuenta su dirección. En este caso, el error promedio absoluto entre los valores predichos y los valores reales del precio del coche es de 0.054 unidades. Este valor es bastante bajo, lo que indica que las predicciones del modelo son, en general, bastante precisas.

3. Mean Squared Error (MSE): 0.005

El MSE es la media de los cuadrados de los errores, y penaliza los errores grandes más que los pequeños debido a la elevación al cuadrado. Un MSE de 0.005 indica que, en promedio, los errores en las predicciones son muy pequeños, lo que sugiere que el modelo predice el precio del coche con un alto grado de precisión.

4. Root Mean Squared Error (RMSE): 0.069

El RMSE es la raíz cuadrada del MSE y proporciona una interpretación más intuitiva del error en las mismas unidades que la variable dependiente. En este

caso, el RMSE de 0.069 indica que la desviación estándar promedio de los errores de predicción es baja, lo que refuerza la precisión del modelo.

5. Mean Signed Difference: 0.021

Este valor refleja la media de las diferencias entre los valores predichos y los reales, con signo. Un valor positivo sugiere que, en promedio, el modelo tiende a sobreestimar ligeramente el precio del coche en 0.021 unidades.

6. Mean Absolute Percentage Error (MAPE): 0.507

El MAPE mide el error absoluto promedio en términos porcentuales. Un MAPE de 0.507 indica que, en promedio, las predicciones del modelo difieren en un 50.7% del valor real del precio del coche. Aunque este porcentaje podría parecer alto, es importante considerar el contexto y la escala de los valores de la variable dependiente. Se mantiene el valor del primer cálculo ya que en el segundo cálculo no se calculó, no se determina aún la razón

7. Adjusted R^2 : 0.898

El R^2 ajustado es una versión modificada del R^2 que tiene en cuenta el número de predictores en el modelo. Un valor de 0.898 indica que el modelo sigue explicando bien la variabilidad en el precio del coche, incluso después de ajustar por el número de variables predictoras utilizadas.

En conjunto, estas métricas sugieren que el modelo de regresión lineal múltiple desarrollado es bastante robusto y tiene un buen rendimiento en términos de precisión predictiva y capacidad para explicar la variabilidad en los precios de los coches. Sin embargo, el MAPE relativamente alto podría indicar la necesidad de revisar las escalas de las variables o considerar modelos adicionales para mejorar la precisión en ciertos casos específicos.

Desarrollo de Plan Estratégico

Objetivos Medibles del Uso de esta Tecnología

Algunos de los objetivos iniciales que hemos planteado para esta solución son (pero no están limitados a):

- Lograr una precisión del 85% en las predicciones de precios en comparación con los precios reales del mercado.
- Incrementar la competitividad de Foreign Cars en el mercado mexicano en un 15% dentro de los primeros seis meses de implementación.
- Reducir el tiempo de ajuste de precios en un 30% mediante el uso del modelo predictivo.

Integración del Modelo de IA al Proceso Actual de la Empresa

El modelo de IA se integrará en el proceso de toma de decisiones de Foreign Cars a través de un sistema de recomendación de precios automatizado. Este sistema será accesible para los equipos de ventas y marketing, permitiéndoles ajustar los precios en tiempo real basándose en las predicciones del modelo.

Resultados del Algoritmo Desarrollado

Se espera que el algoritmo permita a Foreign Cars establecer precios competitivos y alineados con las expectativas del mercado, mejorando así su penetración y éxito en México.

Posibles Obstáculos y su Solución

- **Variabilidad del Mercado:** La fluctuación de precios en el mercado automotriz podría afectar la precisión del modelo. **Solución:** Actualizar el modelo regularmente con datos nuevos y comparables para mantener la precisión y la capacidad de predicción del modelo.
- **Integración Técnica:** La integración del modelo en los sistemas actuales de la empresa podría presentar desafíos técnicos. **Solución:** Colaborar estrechamente con el equipo de Foreign Cars para asegurar una integración fluida.

Cronograma de Implementación con Hitos

Este proyecto está pensado para ser desarrollado en 6 meses, con su sucesivo seguimiento.

A continuación, se presenta un cronograma inicial, sujeto a cambios y ajustes.

- **Mes 1-2:** Recolección de datos y desarrollo del modelo de regresión.
- **Mes 3:** Validación del modelo y ajustes.
- **Mes 4:** Integración del modelo en los sistemas de Foreign Cars.

- **Mes 5:** Capacitación del personal y lanzamiento del sistema.
- **Mes 6:** Monitoreo y ajustes post-implementación.

El seguimiento del proyecto mediante un diagrama de Gantt se puede realizar en, tiempo real, en el siguiente [enlace](#)

