v. 3, n. 18, 2023

Journal of
Engineering
Research

MACHINE LEARNING APLICADO A LA ELECCIÓN DE MODO CUANDO SE PENALIZA LOS VIAJES POR TIEMPO

Andrés Pava Restrepo
Universidad EIA - Escuela de Ingeniería y Ciencias Básicas, Medellín, Colombia
ORCID: 0000-0001-5764-8869

Iván Reinaldo Sarmiento Ordosgoitia
Universidad Nacional de Colombia – Facultad de Minas, Medellín, Colombia
ORCID: 0000-0001-7287-4573

Juan Diego Pineda Jaramillo
University of Luxembourg, Luxemburgo,
ORCID: 0000-0002-4657-7521

Resumen: La planificación de transporte requiere modelar la elección del modo de viaje para predecir la demanda y comprender las variables causales. Esta elección modal depende de una gran cantidad de variables, por lo que frecuentemente se han utilizado modelos de elección discreta, donde los modos de viaje son alternativas mutuamente excluyentes y conjuntas dentro de diversos marcos. El modelo Logit Multinomial (MNL) es el modelo de elección discreta más utilizado. Sin embargo, el MNL tiene varias limitaciones al suponer que la probabilidad de cada alternativa es independiente de las características del resto de las alternativas. Por otro lado, hay evidencia de que los algoritmos de aprendizaje automático o Machine Learning (ML) funcionan óptimamente para los enfoques estadísticos utilizados para la modelación de elección de modo de viaje, debido a que estos no hacen suposiciones drásticas sobre los datos estudiados y aprenden a representar relaciones no lineales y, en general complejas. Por otro lado, la estrategia de aplicar una racionalización del espacio vial disponible para el vehículo privado busca que los conductores de auto privado elijan otras alternativas para acceder a los centros urbanos debido a la penalización en tiempo que estos perciben. Con el fin de identificar el efecto de la demora en el tiempo de viaje causado por la restricción del espacio vial disponible para el vehículo privado en el centro urbano de la ciudad de Medellín-Colombia, se ha desarrollado un modelo de elección de modo de viaje a través de modelos de ML, permitiendo establecer con gran precisión la demanda al aplicar una política de gestión de tráfico limitando el espacio vial para el vehículo privado. Este enfoque demuestra el gran potencial que tienen los modelos de ML para predecir la elección modal, como una alternativa a los modelos de elección discreta.
Palabras clave: Gestión de la demanda, modelos de demanda de transporte, algoritmos de Machine Learning.

INTRODUCCIÓN

Las ciudades, sobre todo las de más de un millón de habitantes, en general presentan altos niveles de congestión en sus centros urbanos por ser las zonas de mayor atracción de viajes. Medellín, con 2,5 millones de habitantes en el centro de un área metropolitana de casi 4 millones de habitantes, es la segunda ciudad en población y producción económica en Colombia, y presenta unos altos niveles de congestión en el centro urbano, por lo que es urgente la necesidad de analizar diferentes alternativas para reducirlos, una de las estrategias es la aplicación de la racionalización del espacio vial disponible para el vehículo privado, con el fin de que este experimente una demora en sus desplazamientos llevando a los conductores del auto privado a elegir otras alternativas para acceder al centro.

Para lograr identificar el impacto de la racionalización en la demanda de viajes en vehículos privados como estrategia de planificación del transporte, se requiere conocer el comportamiento y decisiones de los usuarios, ya que estos eligen la manera de realizar sus desplazamientos de acuerdo a factores individuales, domésticos y exógenos de cada individuo. Para entender estos comportamientos e identificar las variables que influyen en la elección de modo de viaje, se han utilizados diversos modelos de elección de viajes como los modelos de elección discreta, donde los modos disponibles son alternativas mutuamente excluyentes y a la vez conjuntas dentro de diversos contextos y condiciones (Ben-Akiva, 1985). Hasta finales del siglo 20 el modelo Logit Multinomial (MNL) fue uno de los modelos de elección discreta más utilizado en la planificación de transporte (Rich, 2009), este considera el principio de maximización de la utilidad del usuario y tiene un marco

matemático que permite la estimación de factores que influyen en la elección del modo; sin embargo este tiene limitaciones ya que supone que la probabilidad de elección de cada modo de viaje, es independiente de las características particulares de las otras alternativas (Ortúzar, 2003). Aunque las limitaciones del MNL fueron mejoradas por la introducción de los modelos Mixed Logit y la inclusión de variables latentes, es importante explorar otras alternativas de modelación.

Los algoritmos Machine Learning ML surgen como una alternativa para el análisis de diversos temas en la planificación de transporte, y se ha demostrado que funcionan bien para los enfoques estadísticos utilizados en la modelación de elección de modos de viaje. La ventaja de estos algoritmos es que permiten desarrollar relaciones más complejas basados en los datos conjuntos que caracterizan los diversos usuarios y las decisiones que toman frente a las alternativas ofrecidas y sus condiciones (Pineda-Jaramillo, 2019). El proceso metodológico permite aplicar diferentes algoritmos de decisión de manera supervisada buscando de acuerdo a la estructura de datos estimar el modelo que mejor precisión en la predicción desarrolle en el proceso de aprendizaje.

Así pues, para conocer la elección de los usuarios entre los modos disponibles para acceder al centro de Medellín, este artículo propone desarrollar una serie de modelos de Machine Learning para evaluar el efecto de la medida de restricción del espacio vial en la demanda de viajes hacia el centro urbano entre los usuarios habituales del automóvil. Estos modelos permiten incorporar a los usuarios lexicográficos (numerosos entre los usuarios de auto en el contexto de las ciudades colombianas), entre otras variables que puedan estar fuertemente correlacionadas dentro de las características individuales y las condiciones de los escenarios de elección, sin afectar la precisión en la predicción del modelo.

METODOLOGÍA Y PROCEDIMIENTOS

Para desarrollar el modelo para la elección modal de los usuarios del auto bajo una serie de escenarios hipotético de racionalización del espacio vial, buscando identificar si es posible reducir los altos niveles de congestión en el centro urbano, se llevó a cabo la metodología presentada en la Figura 1. A continuación se describen los pasos de una manera más detallada (Abduljabbar, 2019).

TOMA DE INFORMACIÓN:

Con el fin de identificar la sensibilidad de los usuarios del vehículo particular ante una penalización con tiempo a través de la limitación del espacio vial disponible, se aplicó la técnica de preferencias declaradas (PD) donde se presentaron nueve escenarios para la elección de modo de viaje entre el sistema de transporte público masivo incluyendo Metro (SITVA), transporte público (BUS), y vehículo particular, los cuales presentaban atributos relacionados al costo y el tiempo de viaje para ingresar al centro urbano, los tres modos disponibles en la PD corresponden a los usados para ingresar al anillo del centro de Medellín delimitado por las principales vías que lo circundan.

Como resultado de la técnica se obtuvieron 1818 datos de usuarios del vehículo particular de la muestra total a los cuales se les aplica la restricción debido a la racionalización del espacio vial, asociada al número de viajes diarios realizados hacia el centro de la ciudad desde las diferentes zonas de la ciudad según la última encuesta origen destino de hogares del Valle de Aburrá (AMVA, Área Metropolitana del Valle de Aburrá, 2017), involucrando características socioeconómicas del individuo, características de viaje, y las condiciones de

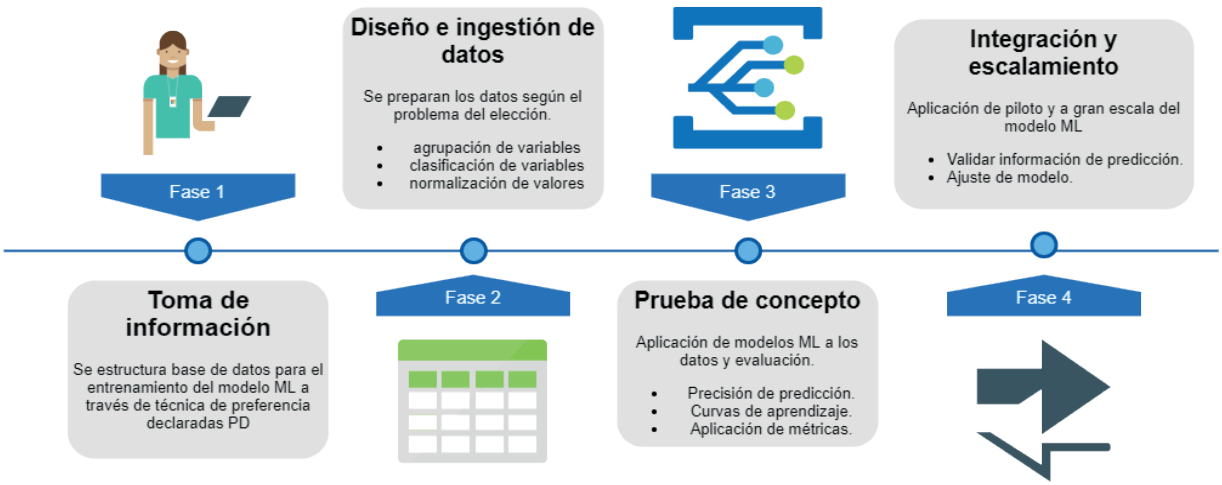

Figura 1. Metodología para el desarrollo de modelos de elección a través de ML

Fuente: Elaboración propia

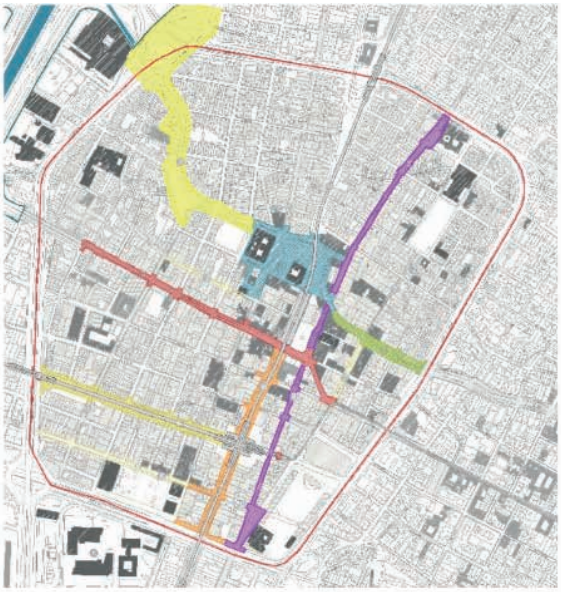

Figura 2. Perímetro de la zona centro de la ciudad de Medellín

Figura 3. Modelo microscópico de la zona centro de la ciudad de Medellín

Fuente: Elaboración propia

	choice	costo_sitva	costo_bus	costo_auto	tv_sitva	tv_bus	tv_auto	gen	ocu	edu	edad	est	mot	frec
0	2	2000	2400	3500	55	30	50	1	0	1	3	2	0	1
1	0	2000	2100	5000	45	40	50	1	0	1	3	2	0	1
2	0	2000	2400	3500	45	50	35	1	0	1	3	2	0	1
3	0	2200	2100	8000	25	30	50	1	0	1	3	2	0	1
4	0	2200	4200	8000	25	40	20	1	0	1	3	2	0	1

Tabla N°1. Estructura de base de datos para cada modo de viaje y sus variables independientes (se presentan solo las primeras 5 filas de los 1818 datos)

Fuente: Elaboración propia

los escenarios de elección de modo de viaje; las variaciones de tiempo frente a la demora se obtienen a partir de curvas volumen demora en los principales corredores viales al interior del centro urbano, construidas a partir del tiempo de viaje obtenido de un modelo microscópico, aplicando la reducción del espacio vial disponible para el auto y el aumento del volumen de circulación en la zona analizada.

DISEÑO E INGESTIÓN DE DATOS:

En esta fase se realiza la preparación de datos, el análisis exploratorio y la extracción de características, las fuentes de datos deben evaluarse y ordenarse según el problema a resolver. En este caso se revisa la distribución de las opciones para cada variable tenida en cuenta para el desarrollo del modelo, en primera instancia se buscan agrupar las características que tienen menor porcentaje de elección de tal manera que el grupo de datos tenga un relativa homogeneidad o un peso significativo dentro del grupo de opciones para la misma variable, esta agrupación debe ser coherente con lo que representa, por ejemplo: si el estrato socioeconómico de menor peso es estrato 1 y estrato 6, no se agrupan ya que no representan la misma condición, sin embargo para la variable nivel de formación bachiller y tecnólogo que tienen menor porcentaje de representación, estos se agrupan ya que son condiciones cercanas. Por otro lado, se hace la identificación de las tipologías de variables utilizadas, es decir para las categóricas consideradas "dummy", y las numéricas, desde los desarrollos ML se puede atribuir esta clasificación con el fin que el modelo identifique claramente los valores de la base de datos según su tipología.

PRUEBA DE CONCEPTO:

Esta fase realiza la aplicación de diferentes modelos de ML de clasificación y su posterior evaluación. Es necesario elegir el modelo que mejor se ajuste a los datos; es importante tener en cuenta las diferencias básicas para lograr una buena validación con base en distintas métricas de precisión para los modelos probados. En este caso se aplicaron 7 tipos de familias de modelos de clasificación con diversas combinaciones de hiperparámetros, que mejor adaptación tuviera a los datos obtenidos de la PD, validando su comportamiento utilizando el método estratificado de validación cruzada 10 veces, ya que es la metodología más aceptada para evaluar un modelo de clasificación de una base de datos (Witter, 2005). Para los modelos con mejor nivel de predicción se aplican métricas de evaluación como el índice Jaccard, el F1 score y la matriz de confusión, además de identificar si existe sobre-aprendizaje, refiriéndose a que los modelo fallan en reconocer un dato de un nuevo individuo, porque este no tiene exactamente los mismos valores de los individuos con los que se entrenó, es decir el modelo no es capaz de generalizar conceptos.

INTEGRACIÓN Y ESCALAMIENTO:

En esta fase se desarrolla el piloto inicial e implementación a gran escala, se valida la información de predicción. Es importante resaltar que el modelo siempre puede ajustarse con base en nueva información recolectada para lograr unos resultados más precisos. Para este caso los modelos se entrenaron con el 75% de los datos obtenidos de las PD, y posteriormente se validaron con el 25%.

RESULTADOS Y DISCUSIÓN

Previamente a la aplicación de los modelos de Machine Learning, fue necesario realizar una limpieza y procesamiento de los datos, que incluía, la categorización de las variables entre numéricas y categóricas, la normalización de variables, verificación de correlación entre

variables, división de los datos aleatoriamente entre datos de entrenamiento y datos de prueba, entre otros. Todo este procesamiento se realizó utilizando diferentes librerías de Python 3.7, como scikit-learn, pandas, matplotlib, numpy. Por otro lado, se realiza el análisis estadístico de los resultados de las encuestas PD de un total de 701 encuestados, de los cuales los usuarios del auto son 202, para 9 escenarios de elección, con un total de 1818 datos, este análisis con el fin de realizar las correspondientes agrupaciones

Una vez procesados los datos, se realizó el entrenamiento de diferentes tipos de modelos de ML, ajustando sus hiperparámetros y aplicando la metodología de validación cruzada (Payam Refaeilzadeh, 2008) técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba, y así encontrar el modelo más preciso. La tabla 2 y figura 2 presenta la precisión media por tipo de modelo de Machine Learning probado con el número de combinaciones resultante entre las diferentes combinaciones de hiperparámetros y evaluaciones aplicando validación cruzada aplicada a cada tipo de modelo.

Las precisiones se encuentran entre 60.3% y 71.3% para la predicción con el grupo de datos utilizado para entrenar los modelos, el clasificador de aumento de gradiente (GB) logro mejor resultado, por lo tanto, se verificar el modelo de elección. Para la validación inicial, se utilizan específicamente dos métricas: la curva de aprendizaje (figura 5) y la matriz de confusión (figura 6). Al evidenciar que el modelo no predice bien para el choice=0 (SITVA) y el choice=1 (BUS), además, al no evidenciar una convergencia en la curva de aprendizaje, se revisa si la muestra presenta un desbalance en las clases que se predicen (el CHOICE) modo de viaje. Al revisar la composición se evidencia que el 14.8% de las encuestas tienen como CHOICE el BUS, el 33.0% con CHOICE el SITVA, y el 52.2% con CHOICE el AUTO. Así pues, la muestra se encuentra desbalanceada y por eso el modelo presenta un mayor nivel de precisión para predecir el modo AUTO. Para eliminar el desbalanceo utilizamos la técnica del sobre muestreo a las clases minoritarias (SITVA y BUS), aumentando aleatoriamente el número de instancias para estas clases. Posterior a este sobre muestreo se vuelve a probar el modelo GB seleccionado, obteniendo una mejor curva de aprendizaje (figura 7) y matriz de confusión (figura 8).

También se evalúa el indicador Jaccard (que estima si el conjunto de opciones de elección pronosticadas coincide estrictamente con el conjunto real de datos) obteniendo un valor de 0.82 (cerca al 1.0 ideal). Adicional, se evalúa también el modelo utilizando el F1-score, obteniendo unos valores de 82% para el SITVA, 85% para el bus y 78% para el auto.

Finalmente se concluye que el modelo Gradient Boosting Classifier (BG) con los parámetros learning_rate = 0.1, max_depth = 3, n_estimators = 1000, subsample = 1.0) es confiable para predecir la elección modal para evaluar la racionalización del espacio vial disponible para el vehículo privado.

Para predecir la elección modal de los usuarios de auto, se realizó un agrupamiento de los diferentes usuarios "típicos" encontrados y aplicar el modelo de ML, y conocer su flexibilidad a cambiar su modo de transporte si el costo y el tiempo de viaje de auto se modificaran dentro de los rangos evaluados en las encuestas PD. Es importante aclarar que este método de buscar los usuarios "típicos" es usado pensando en la "aplicación posterior" a una muestra generalizada de la población, por lo que no significa que se busquen los usuarios más comunes dentro de la base de datos. Para esto se aplicó el modelo K-Means, el cual es un modelo de ML no-supervisado utilizado para

Modelo (ML)	Combinaciones	Precisión media del tipo de modelo
Regresión logística (LR)	9000	0.607
Ridge Classifier (RC)	6000	0.602
vecinos próximos (KNN)	1800	0.604
Máquinas de Vector de Soporte (SVC)	12000	0.635
Bagged Decision Trees (BDT)	17280	0.613
Random Forest (RF)	3600	0.603
Gradient Boosting (GB)	2430	0.712

Tabla N°2. Resultados de precisión de la predicción de los modelos con los datos de aprendizaje

Fuente: Elaboración propia Python

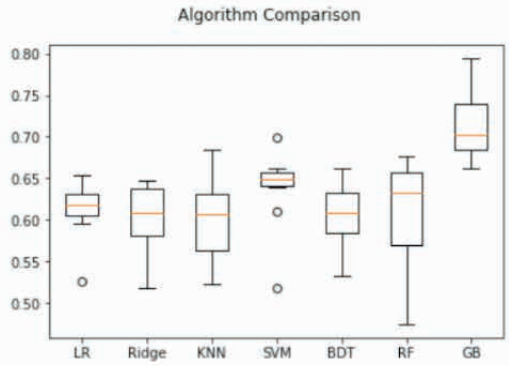

Figura N°4. Comparación de la precisión de los algoritmos (ML)

Fuente: Elaboración propia Python

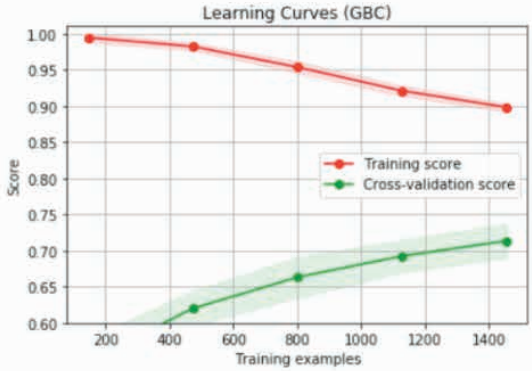

Figura N°5. Curvas de aprendizaje para GBC

Fuente: Elaboración propia

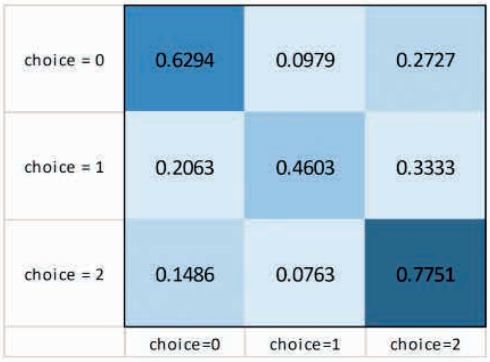

Figura N°6. Matriz de confusión para los modos de viaje

Fuente: Elaboración propia

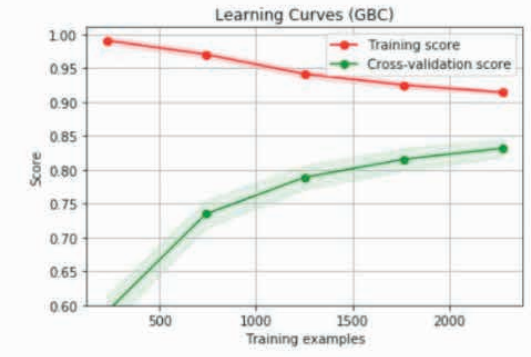

Figura N°7 Curvas de aprendizaje para GB datos balanceados

Fuente: Elaboración propia Python

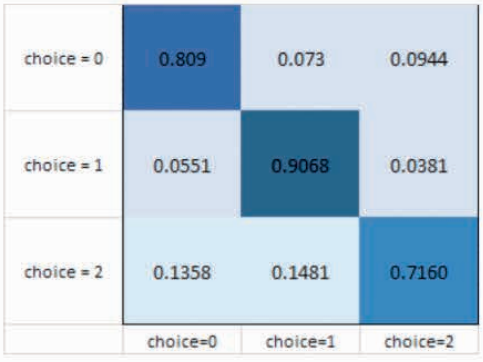

Figura N°8. Matriz de confusión para los modos de viaje

Fuente: Elaboración propia Python

simplificar grandes bases de datos por medio de agrupar variables con valores similares en pequeños números de categorías homogéneas (Pineda, 2019); el agrupamiento se hace a partir de las variables categóricas dejando por fuera el costo y el tiempo. Este modelo separa la muestra en un número de grupos de tal que se minimice la suma de las distancias cuadradas entre cada par de puntos de cada grupo, se utiliza el "método del codo" para definir el número óptimo de grupos a dividir la muestra. Como resultado, se optó por elegir un valor de k= 6 y 7 (Figura 9), definiendo seis grupos de usuarios "típicos".

Considerando que estos grupos de usuarios "típicos" configuran el total de usuarios del auto de la muestra total, es necesario configurar la curva de probabilidad de elección de modo auto de manera generalizada, con la distribución de los grupos de usuarios de la tabla 3 y su probabilidad de elección. La curva de probabilidad de elección general resultante se correlaciona con una curva de tendencia que representa la variación de la demanda de acuerdo al tiempo de viaje debido a la demora como se ve en la Figura 9.

CONCLUSIONES

Para el desarrollo de modelo de demanda de los usuarios de autos se pueden usar modelos de elección discreta o hacerlo a partir de modelos de Machine Learning (ML), para su uso es necesario que el número de datos sea grande, por lo tanto, se requiere un número de encuestas alto (mayor a 200 individuos), incluso si la muestra necesaria para el nivel de confiabilidad requerida sea menor (por ejemplo los 100 mínimos que arrojan las formulas estadísticas convencionales), la causa es que en la medida que existan más datos para el aprendizaje del modelo este puede predecir mucho mejor y sus resultados pueden ser más confiables. De igual manera para el caso de las variables objetivos (choice), entre más homogéneo sea la proporción de los resultados de elección modo, la precisión de predicción para cada uno será similar y mejor, pero en la realidad esta proporción homogénea es difícil, ya que es una elección aleatoria, lo que refuerza aún más la necesidad de trabajar con un mayor número de datos para el desarrollo de modelos de ML.

En el caso de que los datos obtenidos para el desarrollo de un modelo de predicción de modo sean pocos, puede ser recomendable utilizar modelos de elección discreta MNL. Si bien muchos autores han demostrado que los modelos de ML predicen mejor, este depende del número de datos disponible para el aprendizaje, en cambio los modelos MNL simplifican el manejo de los datos ya que no desarrolla relaciones tan complejas como las realizadas por los modelos de ML y por lo tanto la precisión de predicción podría llegar a ser mejor. Para este caso se llegó a una buena precisión del 71% en la elección de los automovilistas para el modelo ML.

Es normal que en la realidad las variables objetivo estén desbalanceadas, ya que dependiendo del grupo de encuestados y las características de la población a encuestar el resultado de elección en las PD puede estar más orientado a un modo que a otros, además de la distribución socioeconómica de la población en general. Por esto es necesario aplicar balanceo en los datos con el fin que el número utilizado para el aprendizaje sea igual y suficiente para una buena precisión en la predicción de cada modo.

Aun así, con el 71% de precisión en el modelo de ML utilizado, se evidencia que el tiempo de viaje afecta la decisión de viaje de los usuarios de auto, y se puede construir una curva de demanda que relaciona el tiempo de viaje con la cantidad de viajes en auto realizados. Si bien no se percibe su costo en la misma proporción que un impuesto económico, los individuos evitan elegir el auto

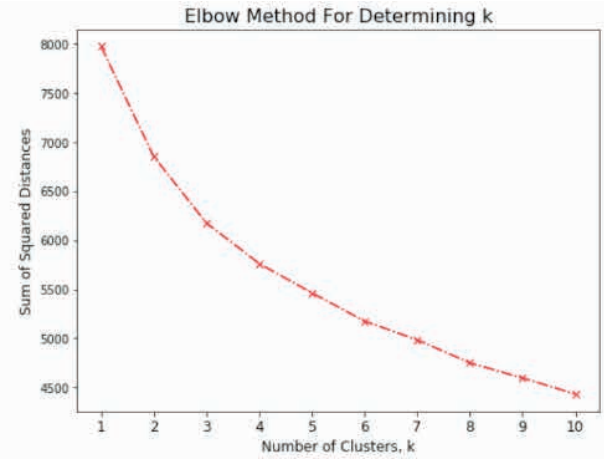

Figura 9. Curva de aplicación para determinar el número de grupos K

Fuente: Elaboración propia

Grupo	*Total, de datos agrupados*	*Porcentaje del total*
Cluster 1	216	11.87%
Cluster 2	342	18.81%
Cluster 3	243	13.36%
Cluster 4	378	20.78%
Cluster 5	297	16.37%
Cluster 6	342	18.81%
Total	**1818**	**100%**

Tabla 3. Distribución de datos agrupados por clúster de "usuarios típicos"

Fuente: Elaboración propia

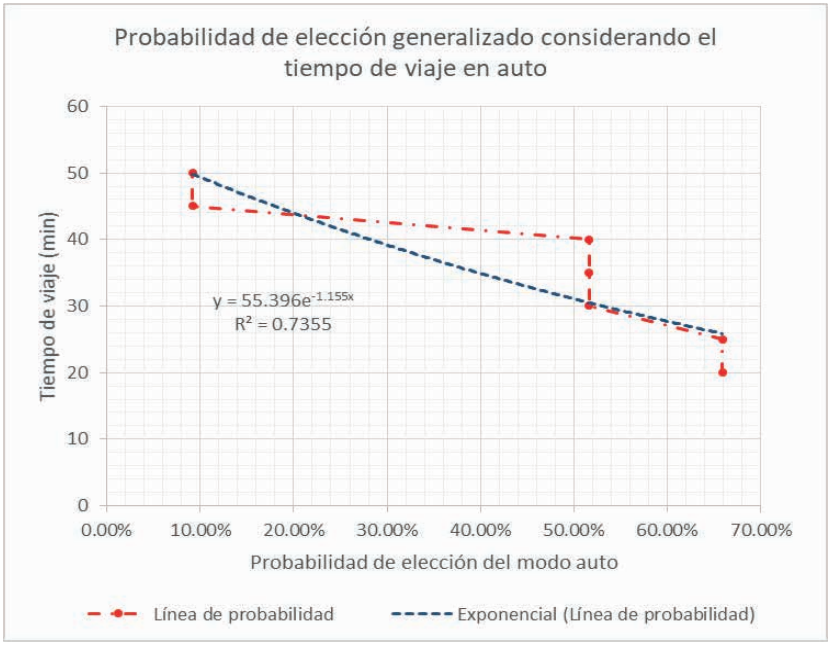

Figura 9. Probabilidad de elección generalizada del modo auto, de acuerdo con la variación del tiempo de viaje.

Fuente: Elaboración propia

en la medida que el tiempo de desplazamiento aumenta.

REFERENCIAS

Abduljabbar, R. D. (2019). Applications of artificial intelligence in transport: an overview. *Sustainability, 11(1)*, 189-190.

AMVA, Área Metropolitana del Valle de Aburrá. (2017). *Encuesta Oringen - Destino de Hogares.* Medellín. Obtenido de https://eee.metropol.gov.co/encuesta_od2017_v2/index.html#/

Ben-Akiva, M. a. (1985). Discrete choice analysis: theory and application to travel demand. *MIT Press, Boston*.

Ortúzar, J. y. (2003). El problema de modelación de demanda desde una perspectiva desagregada: el caso del transporte. *Eure*, 149-171.

Payam Refaeilzadeh, L. T. (2008). k-fold Cross-Validation. 1-6.

Pineda, J. (2019). A review of Machine Learning (ML) algorithms used for modeling travel mode choice. *DYNA 86(211)*, 32-41.

Pineda-Jaramillo, J. (2019). A review of Machine Learning (ML) algorithms used for modeling travel mode choice•. *DYNA, 86(211)*, 32-41.

Rich, J. H. (2009). A weighted logit freight mode-choice model. *Journal of Transportation Research Part A: Policy and Practice*, 1006-1019.

Witter, I. E. (2005). *Practical machine learning tools and techniques.* Morgan Kaufmann.