Técnicas de Machine Learning en el estudio de viajes realizados por los habitantes del Valle de Aburrá

UNIVERSIDAD DE ANTIOQUIA
Facultad de Ingeniería

Angie Paola Correa Sepúlveda & Olga Cecilia Úsuga Manco
Departamento de Ingeniería Industrial, Facultad de Ingeniería, Universidad de Antioquia
angie.correa@udea.edu.co; olga.usuga@udea.edu.co

Resumen

En este trabajo se desarrolló un análisis descriptivo y predictivo del modo de transporte elegido por los habitantes del Valle de Aburrá a partir de variables como el municipio de origen, comuna de origen, hora de inicio del viaje, municipio de destino, comuna de destino, motivo del viaje, entre otras características individuales, utilizando metodologías de machine learning como la regresión logística multinomial, bosques aleatorios, máquinas de vectores de soporte, redes neuronales y K-vecinos más cercanos. Se encontró que el modelo de bosques aleatorios tiene un mejor desempeño al presentar una mejor tasa de clasificación correcta del modo de transporte, incluso mayor que el modelo de regresión logística multinomial que es el que tradicionalmente se usa. Finalmente, se construyó una aplicación web utilizando el paquete Shiny de R en donde los usuarios podrán visualizar e interactuar con los datos usados en el análisis.

Introducción

La elección del medio de transporte de los habitantes ha sido un aspecto importante a estudiar para el desarrollo de proyectos relacionados con la movilidad e infraestructura [1]. Modelos tradicionales como la regresión logística multinomial han sido ampliamente utilizados para la predicción de los modos de transporte [2] debido a la facilidad en su aplicación e interpretación de resultados. En este trabajo se presenta la aplicación de diferentes técnicas de Machine Learning para la predicción del medio de transporte elegido por los habitantes del Valle de Aburrá y se compara la eficiencia de clasificación de las diferentes técnicas usadas.

Metodología

Los datos analizados corresponden a la Encuesta Origen Destino 2017 realizada por el Área Metropolitana a algunos habitantes de los diez municipios que comprenden el área metropolitana del Valle de Aburrá. Para la aplicación de las técnicas de Machine Learning se tuvieron en cuenta las siguientes variables:

- Variables propias del individuo: edad, género, escolaridad, ocupación e ingresos.
- Variables del viaje realizado: municipio de origen, comuna de origen, hora de inicio del viaje, comuna de destino, duración del viaje, motivo del viaje y modo de transporte.

Se ajustaron algunos modelos de clasificación de aprendizaje automático como la regresión logística multinomial, bosques aleatorios, máquinas de soporte vectorial, k-vecinos más cercanos y redes neuronales, con el objetivo de predecir el modo de transporte de los habitantes y se compararon entre sí utilizando el porcentaje de clasificación correcta. Todo lo anterior se realizó en el lenguaje de programación R [3].

Conclusiones

- El modelo de bosques aleatorios resultó ser el más preciso al registrar la mayor tasa de clasificación correcta del modo de transporte, incluso por encima del modelo de regresión logística multinomial, el cual es el tradicionalmente utilizado.
- La variable con mayor importancia para el transporte público fue la duración del viaje, mientras que para el modo de transporte “moto” fue la edad.
- En general, la duración del viaje es un factor importante a la hora de elegir el medio de transporte, seguido por la edad, mientras que la hora de origen no fue un factor clave.

Análisis exploratorio

En total se analizaron 7529 viajes distribuidos sobre el Valle de Aburrá. En la Figura 1 se muestra el flujo de viajes entre los diez municipios que comprenden el área metropolitana, siendo el municipio de Medellín el de mayor frecuencia de destino con un 52.9% del total de datos analizados, mientras que el municipio de destino con menor frecuencia es La Estrella con apenas un 0.89%.

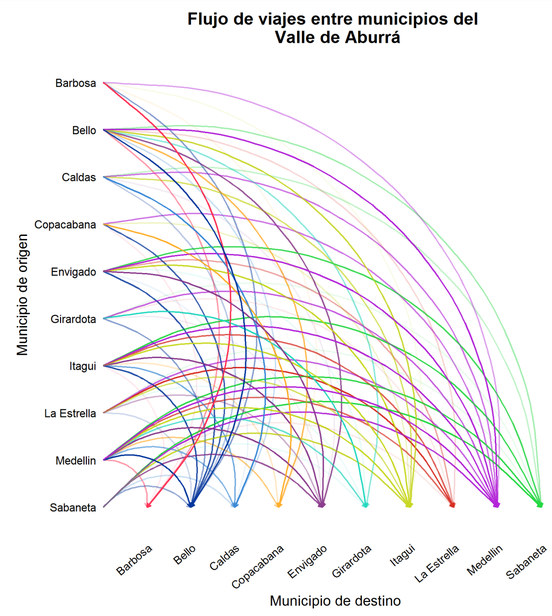

Figura 1. Flujo de viajes en el Valle de Aburrá.
Fuente: elaboración propia.

También se analizó el flujo de viajes durante el día, y como es de esperarse, el mayor flujo ocurre en horas de la mañana y finalizando la tarde, así que la duración total del viaje suele aumentar como se observa en la Figura 2. Un 26.3% de las personas se desplazan entre las 6 a.m. y las 8 a.m., mientras que el 24.2% lo hace entre las 4 p.m. y 5 p.m., especialmente como retorno a sus hogares luego de la jornada de trabajo.

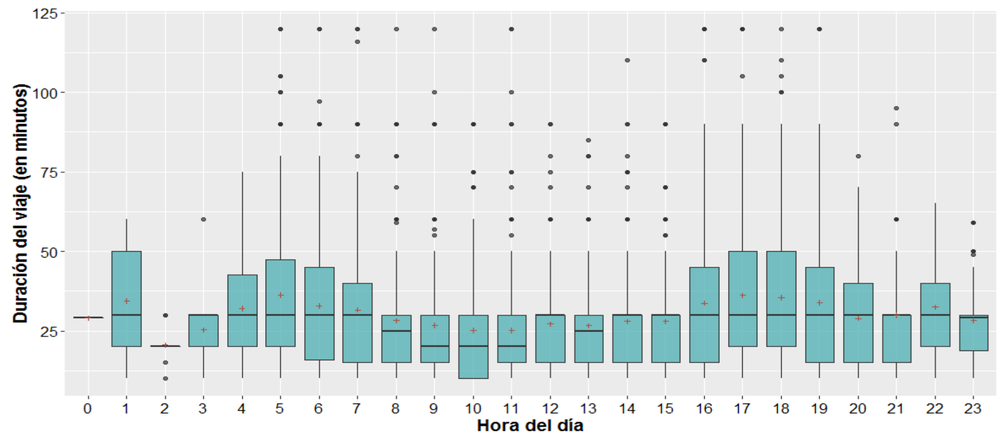

Figura 2. Duración del viaje en función de las horas del día.
Fuente: elaboración propia

Predicción del modo de transporte

Se utilizó la matriz de confusión para evaluar la eficiencia de clasificación de cada modelo y los resultados se muestran en la Tabla 1. Los modelos con mejor tasa de clasificación fueron los bosques aleatorios, en el cual se utilizaron 1000 árboles, y las máquinas de soporte vectorial, utilizando la función Kernel radial. Estos dos modelos tuvieron mejor eficiencia que la regresión logística multinomial, que es el modelo tradicionalmente utilizado para estos casos.

Tabla 1. Tasa de clasificación correcta para cada modelo.

Modelo	Precisión de clasificación
Regresión logística multinomial	0.622
Bosques aleatorios	0.651
Máquinas de soporte vectorial	0.624
Redes neuronales	0.607
K-vecinos más cercanos	0.612

Referencias

[1] de Dios Ortuzar, J., & Willumsen, L. G. (2011). Modelling transport. John Wiley & Sons.
[2] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in econometrics. New York, NY: Academic Press.
[3] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.