**Problem Set 4: Predicting Tweets**

1. **Introducción**

En los últimos años, se ha generado un especial interés por parte de la academia para realizar análisis técnicos sobre las redes sociales que conduzcan a generar evidencia principalmente en época de comicios, para la toma de decisiones de los partidos políticos, candidatos y ciudadanos. Caso puntual objeto de análisis es la red social Twitter, la cual se ha convertido en una de las más empleadas en el mundo, al ser un espacio que permite a las personas comunicar y estar en contacto a través de mensajes rápidos y frecuentes. Los Tweets que pueden contener fotos, videos, enlaces y texto (Eskibel, 2022), y especialmente por las siguientes características (Gomila, 2020):

1). Es una plataforma que favorece la viralidad de la información publicada en comparación con otras redes sociales, dado su diseño, interfaz y disponibilidad de recursos (hashtags, retuits y etiquetas); 2). Es la red social más utilizada por el sistema político y el mediático. Es decir, a pesar de no ser la plataforma más empleada, es la que tiene el público más influyente; 3). Twitter se ha convertido en una herramienta fundamental para el diseño y la ejecución de la estrategia comunicativa en campaña electoral de los partidos políticos y; 4). Su estructura facilita a los investigadores un acceso más directo a los datos, simplificando su extracción, procesamiento y análisis (Gomila, 2020).

Adicionalmente, esta social media cuenta con características que la hacen única a la hora de compararlas con otras. Entre las cuales se destacan: 1). Impacto: Están los líderes de opinión, políticos, gobernantes, candidatos, dirigentes, periodistas, medios de comunicación, referentes sociales y culturales, formadores de opinión, gente activa e influyente; 2). Noticia: Suele ser una [vía más directa para salir en las noticias](https://danieleskibel.com/noticias/) que las tradicionales ruedas de prensa; 3). Contacto: es una forma rápida de contactar con personas relevantes para el político; 4). Brevedad: un tweet son 140 caracteres. No es una columna ni un discurso ni unas largas declaraciones; 5). Velocidad: es una herramienta ideal para difundir o seguir en tiempo real las novedades de un evento que está desarrollándose; 6). Interacción: permite intercambiar ideas, dialogar, discutir, defender posiciones, responder; 7). Síntesis: facilita a los políticos sintetizar su pensamiento en frases breves que sirvan para un título periodístico y que sean fácilmente recordadas; 8). Receptividad. El público de Twitter recibe con mucha mayor naturalidad los mensajes vinculados a la política; 9). Movilidad: Twitter es perfecto para usar desde los smartphones (Eskibel , 2022).

En este sentido, Twitter se ha convertido en un espacio para la difusión de ideas, información y sentimientos por parte de los usuarios. No es la más utilizada a nivel mundial, pero sí es la más influyente en el discurso público. En las últimas décadas se ha convertido en la plataforma para que muchos políticos de todo el mundo se comuniquen con su audiencia y puedan dar a conocer sus iniciativas de política pública y de mandato de gobierno. Ejemplo de ello, se ha visto en las elecciones de Estados Unidos, pues la campaña electoral de Obama en 2008 marcó el camino. Twitter fue una de sus herramientas más novedosas, y el día de la elección ya tenía 100 mil seguidores. En 2012, apenas 4 años después, Obama llegó al día de la elección con 20 millones de seguidores (Eskibel , 2022).

Sin embargo, por la cantidad de cifras que se manejan en las redes sociales, y en especial esta, resulta complejo hacer predicciones sobre sus tweets. Este estudio se basa en un modelo predictivo que contribuye a predecir a quién pertenece cada tuit. El conjunto de datos de capacitación contiene tuits de las cuentas de tres destacados políticos colombianos: Claudia López, Gustavo Petro y Álvaro Uribe. El conjunto de datos de prueba contiene 500 tweets sin etiquetar.

En el presente documento se consideran cuatro (4) modelos predictivos y se profundiza en el que presenta mejor desempeño. Se evalúan una regresión lineal simple, regresiones lineales regularizadas de Lasso y Ridge y, finalmente, un modelo Random Forest. Como resultado del ejercicio, se obtiene que el modelo Random Forest presenta el mejor desempeño en la predicción de precios de vivienda en la localidad de Chapinero, con un RMSE (Root Mean Squared Error) de COP $ 239.012.001,98.

**Nota:** La base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: <https://github.com/Yilap/Repositorio_Taller4.git>

**Contexto**

En Colombia, en el primer semestre de 2022 se vendieron en promedio 127.218 viviendas nuevas, representando un crecimiento de 2,5% en comparación con el mismo periodo del año inmediatamente anterior (Camacol, 2022). Adicionalmente, 7 de cada 10 viviendas vendidas fueron de interés social y las regiones que generaron más ventas fueron: Bogotá con 27 mil unidades, Valle con 18 mil, y Atlántico con 15 mil. Lo anterior representa 27,1 billones de inversión en vivienda en lo corrido del año (Camacol, 2022). Por otro lado, informes de la Cámara Colombiana de Construcción demuestran que el sector inmobiliario terminó el 2022 con un balance positivo, donde las ventas fueron 66.368 viviendas No VIS y 168.224 unidades de VIS (Semana, 2022). El promedio de las ventas ascendió en un 28% con relación a los últimos años y se estima la construcción de alrededor de 360.000 viviendas en construcción para el futuro.

1. **Datos**
2. ***Descripción de las fuentes de datos***

Para el desarrollo de este Problem Set se utilizarán los siguientes datos: por un lado, la tasa de homicidios y tasa de hurto a residencias por UPZ (Unidad de Planeación Zonal), entendida como una división administrativa de la ciudad que es más pequeña que las localidades, pero más grande que los barrios. Tanto la tasa de homicidios como la de hurto a residencias son tomadas de la Secretaría Distrital de Seguridad, Convivencia y Justicia y, las estimaciones de población de las UPZ de 2018 a 2021 se extrajeron de la Secretaría Distrital de Planeación. Se procede a dividir la ocurrencia del crimen entre la población por 100 mil habitantes y se obtiene la tasa por 100 mil habitantes.

1. ***Análisis descriptivo de los datos (estadísticas descriptivas)***

El precio de una vivienda puede estar determinado por diversos factores tanto económicos como sociales. Para el caso de estudio, se analizan factores internos (propios de las viviendas) como externos (elementos geoespaciales como distancia a hospitales, parques y colegios, etc.). En este sentido, es necesario tener información de variables asociadas al precio, con el propósito de obtener un modelo robusto. Para el análisis de los modelos de predicción se utiliza una muestra tomada de Properati sobre los precios de venta y características de inmuebles ubicados en la localidad de Chapinero en Bogotá, lo que arroja una muestra representativa de entrenamiento y testeo del precio de venta de las viviendas en esta zona, así como de los principales atributos que definen el precio de mercado de esta.

1. **Modelos y resultados**
2. **Conclusiones y recomendaciones**
3. **Bibliografía**

Twitter., (s.f.). Preguntas frecuentes para usuarios nuevos. Centro de ayuda. Recuperado de: <https://help.twitter.com/es/resources/new-user-faq>

Eskibel, D., (2022). 10 razones por las que los políticos prefieren Twitter. Recuperado de: <https://danieleskibel.com/twitter10/>.

Gomila, G., (2020). ¿Para qué usan Twitter los partidos en campaña?. Recuperado de: <https://agendapublica.elpais.com/noticia/13745/qu-usan-twitter-partidos-campana>.