Universidad Nacional Autónoma de México

Modelo predictivo para mejorar la recomendación de restaurantes en la aplicación Uber Eats

Introducción a la Ciencia de Datos en Escenarios Médico-Biológicos

Aldo Martínez Cruz

29/11/2024

1 Introducción

1.1 Contexto

En la actualidad, los teléfonos inteligentes (smartphones) se han convertido en herramientas esenciales en la vida diaria. Más allá de ser dispositivos de comunicación, ofrecen múltiples funciones que facilitan actividades como compras, pagos y acceso a servicios. Esto ha llevado a las empresas a desarrollar aplicaciones móviles que sean intuitivas y fáciles de usar, mejorando así la experiencia del cliente y promoviendo una mayor fidelidad. Estas aplicaciones no solo simplifican las tareas cotidianas, sino que también se han transformado en un canal estratégico para incrementar ventas y consolidar la presencia de las empresas en un mercado digital competitivo.

La industria de los servicios de comida a domicilio experimentó un crecimiento sin precedentes durante la pandemia de COVID-19. Esto se debió a la necesidad de las personas de permanecer en casa mientras seguían disfrutando de comidas preparadas. Se proyecta que para 2025 los ingresos de este sector aumenten un 20 %, y para 2029 se espera que el mercado alcance un valor de $1.85 mil millones de dólares. Este crecimiento refleja cambios en los hábitos de consumo y la conveniencia que ofrecen estas plataformas. Sin embargo, las empresas enfrentan retos importantes, como adaptarse a las preferencias cambiantes de los consumidores y competir en un mercado saturado. Para alcanzar los objetivos proyectados, deben implementar estrategias innovadoras que las diferencien y respondan a las demandas de los clientes.

Uber Eats es un servicio internacional de entrega de comida a domicilio a través de una aplicación móvil. Es considerado el líder mundial en este sector, con 88 millones de usuarios y ganancias de $12.1 mil millones de dólares en 2023. Sin embargo, su posición está constantemente amenazada por la competencia de otras empresas en distintas regiones, lo que genera un mercado saturado y altamente competitivo. Este entorno obliga a los usuarios a comparar plataformas y elegir la que mejor se adapte a sus necesidades, considerando aspectos como la variedad de opciones, el costo y la calidad del servicio. Para

mantener su liderazgo, Uber Eats necesita seguir innovando y adaptándose a las dinámicas de este exigente mercado global.

1.2 Justificaión

Los resultados de este proyecto podrían tener un impacto significativo en las ganancias de Uber Eats al optimizar el sistema de recomendaciones de platillos dentro de la aplicación. Un sistema de recomendaciones más preciso y adaptado a los gustos y preferencias individuales de los usuarios permitiría una experiencia más personalizada, lo que podría llevar a un aumento en el número de pedidos y, por lo tanto, en las ganancias de la empresa. Al mejorar la precisión de las recomendaciones, Uber Eats podría incrementar la satisfacción del cliente y fomentar una mayor frecuencia de uso de la aplicación, lo que se traduce en un mayor volumen de ventas.

Por otro lado, los usuarios también se beneficiarían al encontrar una mayor variedad de platillos que se ajusten a sus gustos y preferencias, lo que enriquecería su experiencia al utilizar la aplicación. En lugar de tener que buscar entre una gran cantidad de opciones, podrían recibir sugerencias más relevantes, facilitando su elección y aumentando la probabilidad de realizar un pedido. Esto podría resultar en una mayor satisfacción del cliente y fidelización.

Además, el aumento en las ventas en la aplicación no solo beneficiaría a Uber Eats, sino también a los repartidores, quienes verían un incremento en sus ingresos a medida que se generaran más pedidos. Dado que los repartidores reciben una compensación basada en la cantidad de entregas realizadas, un aumento en las ventas implicaría más trabajo para ellos,

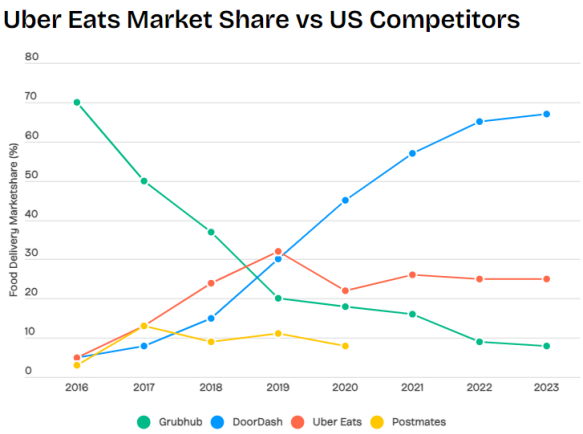

Figura 1: Cuota de mercado de empresas de la industria de servicio online de comida a domicilio.

lo que resultaría en mayores ingresos para quienes forman parte de la red de distribución de Uber Eats.

Mejorar las recomendaciones dentro de la aplicación podría incrementar la cuota de mercado de Uber Eats, un indicador clave que mide el porcentaje del mercado total controlado por la empresa dentro del sector de comida a domicilio. Este porcentaje refleja su capacidad para atraer y retener clientes en comparación con sus competidores. Aumentar la cuota de mercado implica no solo un incremento en las ventas, sino también un fortalecimiento de la presencia de Uber Eats como líder en el sector.

En un mercado saturado y competitivo, donde empresas como DoorDash han superado recientemente a Uber Eats en algunas regiones (Figura 1), optimizar la experiencia del usuario mediante un sistema de recomendaciones más preciso podría ser decisivo para recuperar terreno perdido. Al consolidar su cuota de mercado, Uber Eats no solo aseguraría un crecimiento sostenido, sino que también afianzaría su posición como la opción preferida por los consumidores.

1.3 Problema o necesidad

El problema central de este proyecto radica en la naturaleza dinámica de las preferencias de los usuarios hacia los restaurantes y platillos, las cuales pueden cambiar rápidamente. Este comportamiento volátil plantea un desafío para los sistemas de recomendación, ya que las sugerencias deben adaptarse en tiempo real a las nuevas preferencias de los usuarios para seguir siendo relevantes y útiles. La solución propuesta busca mejorar la calidad de las recomendaciones dentro de la aplicación, asegurando que estas se ajusten constantemente a los cambios en los gustos de los usuarios.

Esta mejora en el sistema de recomendaciones es fundamental para Uber Eats, ya que tiene un impacto directo en la generación de ventas y en la fidelización de los usuarios. Al proporcionar opciones más personalizadas y acertadas, se incrementa la probabilidad de que los usuarios realicen más pedidos, lo que resulta en un aumento en los ingresos de la plataforma. Además, al retener a los usuarios mediante una experiencia más ajustada a sus gustos, Uber Eats puede mantener su popularidad en un mercado competitivo.

Por otro lado, los propios usuarios se benefician al encontrar platillos que se alinean mejor con sus preferencias personales. Esto mejora su experiencia dentro de la aplicación, ya que no tienen que perder tiempo buscando opciones que no les interesan. Al ofrecer recomendaciones más precisas, se facilita la toma de decisiones, lo que aumenta la satisfacción y la probabilidad de que continúen utilizando la plataforma a largo plazo. En resumen, mejorar la calidad de las recomendaciones no solo beneficia a Uber Eats en términos de ingresos, sino que también enriquece la experiencia del usuario, ayudando a mantener su posición como el servicio más popular en la industria.

2 Objetivos

2.1 Objtetivo general

- Realizar un modelo predictivo que mejore las recomendaciones de restaurantes en la aplicación de Uber Eats.

2.2 Objetivos específicos

- Buscar en la base de datos de las ventas de Uber Eats los datos que se utilizarán en el proyecto.
- Realizar un Análisis Exploratorio de los datos.
- Encontrar insights de los datos.
- Hacer una ingeniería de características de los datos
- Proponer una solución y adaptar modelos para esta solución.
- Evaluar los modelos utilizados.

3 Marco de Trabajo

3.1 Alcance del proyecto

Se utilizará una base de datos pública que contiene información sobre las ventas de Uber Eats, la cual fue recuperada desde el siguiente enlace: `https://www.kaggle.com/datasets/ahmedshahriarsakib/uber-eats-usa-restaurants-menus?select=restaurants.csv`. Esta base de datos contiene detalles sobre los menús de restaurantes y las ventas, y será clave para el análisis y la mejora de los sistemas de recomendación de la aplicación. Es importante señalar que esta base de

datos puede presentar algunas limitaciones. En primer lugar, la estructura de los datos puede no ser la más adecuada para un análisis directo, lo que podría requerir procesos de limpieza y reestructuración. Además, dado que los datos pueden no estar actualizados, podrían existir algunas inconsistencias o desajustes con respecto a las condiciones actuales del mercado.

3.2 Resultados esperados

- Modelos predictivos con buen rendimiento que mejoren las recomendaciones de restaurantes en la aplicación. Se espera que los modelos desarrollados puedan predecir de manera efectiva las preferencias de los usuarios, mejorando así las recomendaciones de restaurantes y platillos dentro de la aplicación. Un rendimiento adecuado de estos modelos facilitará que las sugerencias se adapten a las preferencias de los usuarios, aumentando la satisfacción y la probabilidad de que realicen pedidos.

3.3 Beneficiarios e Impacto

La marca Uber Eats se vería beneficiada al mantener su posición como el servicio de comida a domicilio más popular, lo que contribuiría a aumentar su cuota de mercado y consolidar su liderazgo. Además, al mejorar las recomendaciones de platillos y restaurantes, Uber Eats podría incrementar sus ganancias, ya que se espera que los usuarios realicen más pedidos debido a una experiencia más personalizada y eficiente.

Los usuarios también se verían beneficiados, ya que recibirían recomendaciones de platillos que se alinean mejor con sus gustos, lo que mejoraría su experiencia al utilizar la aplicación. Esto no solo facilitaría su toma de decisiones, sino que también podría aumentar la satisfacción general y fomentar la fidelización.

Por último, los repartidores se verían beneficiados indirectamente, ya que un aumento en las ventas generaría más pedidos para entregar. Esto implicaría mayores ingresos para los repartidores, quienes reciben una compensación basada en la cantidad de entregas realizadas. Al haber más pedidos debido a las recomendaciones mejoradas, los repartidores tendrían más oportunidades para generar ingresos.

3.4 Recursos y limitaciones

En este proyecto se utilizarán diversas librerías de Python para facilitar el análisis y modelado de los datos. Las librerías principales serán:

- Pandas: para la manipulación y análisis de datos, incluyendo la carga de la base de datos, limpieza y transformación de los datos.
- Numpy: para realizar cálculos numéricos y operaciones con matrices.
- Matplotlib y Seaborn: para la visualización de los datos, creando gráficos que ayuden a interpretar y presentar los resultados de manera clara.
- sklearn: para la implementación de modelos predictivos, tales como algoritmos de clasificación y regresión.

3.5 Descripción de los datos

La base de datos utilizada proviene de un conjunto de datos público sobre restaurantes registrados en la aplicación de Uber Eats en Estados Unidos. Esta base de datos se recuperó desde el siguiente enlace: `https://www.kaggle.com/dataset`

s/ahmedshahriarsakib/uber-eats-usa-restaurants-menus?select=restaurants.csv, y su archivo se denomina 'restaurants.csv'.

Para tener una primera visión de la estructura de los datos, se puede revisar los primeros 5 registros como se puede apreciar en la Figura 2. Esto permitirá identificar las columnas disponibles, el tipo de información contenida y verificar si existen valores faltantes o irregularidades que deban ser corregidas antes de realizar un análisis más profundo. La inspección inicial de los datos es fundamental para preparar la base de datos para su posterior procesamiento y análisis.

	id	position	name	score	ratings	category	price_range	full_address	zip_code	lat	lng
0	1	19	PJ Fresh (224 Daniel Payne Drive)	NaN	NaN	Burgers, American, Sandwiches	$	224 Daniel Payne Drive, Birmingham, AL, 35207	35207	33.562365	-86.830703
1	2	9	J' ll''z Smoothie-N-Coffee Bar	NaN	NaN	Coffee and Tea, Breakfast and Brunch, Bubble Tea	NaN	1521 Pinson Valley Parkway, Birmingham, AL, 35217	35217	33.583640	-86.773330
2	3	6	Philly Fresh Cheesesteaks (541-B Graymont Ave)	NaN	NaN	American, Cheesesteak, Sandwiches, Alcohol	$	541-B Graymont Ave, Birmingham, AL, 35204	35204	33.509800	-86.854640
3	4	17	Papa Murphy's (1580 Montgomery Highway)	NaN	NaN	Pizza	$	1580 Montgomery Highway, Hoover, AL, 35226	35226	33.404439	-86.806614
4	5	162	Nelson Brothers Cafe (17th St N)	4.7	22.0	Breakfast and Brunch, Burgers, Sandwiches	NaN	314 17th St N, Birmingham, AL, 35203	35203	33.514730	-86.811700

Figura 2: Estructura de la base de datos, donde se pueden observar los nombres de las columnas.

La base de datos tiene las siguientes columnas:

- id: Representa un índice para cada renglón.
- position: Posición del restaurante en los resultados de búsqueda.
- name: Nombre del restaurante.
- score: Calificación del restaurante de 0 a 5.
- category: Categoría del restaurante con base en el tipo de comida que vende.
- price_range: Rango de precios del restaurante ($= Económico, $$= Moderadamente caro, $$$= Caro, $$$$= Muy caro).
- full_address: Dirección completa del restaurante (calle, región (ciudad), estado, código postal).
- zip_code: Código postal del restaurante.
- lat: Latitud de la ubicación del restaurante.
- lng: Longitud de la ubicación del restaurante.

3.6 Limpieza de los datos

De estas variables, desecharemos las columnas `id`, `name`, `category` y `full_address`, ya que para realizar técnicas de visualización no nos ayudarán. Se eliminó la columna `'id'` porque podemos trabajar con el id que proporciona la propia librería de Pandas. `'category'` se eliminó ya que dentro de sus registros se combinan muchas palabras, lo que hace que prácticamente cada restaurante tenga su propia categoría, dificultando poder agrupar los datos por una sola categoría. Y por último, se eliminó `full_address` porque podemos trabajar con las latitudes y longitudes.

Por otro lado, las columnas `'score'`, `'ratings'` y `'price_range'` presentaban valores NaN, que fueron reemplazados por el valor de la moda de cada columna.

Además, se redondeó a un decimal las latitudes y longitudes para poder trabajar por zonas geográficas dentro de las cuales se pueden encontrar varios restaurantes. Si estas variables tienen más decimales, tendríamos longitudes y latitudes exclusivas para cada restaurante y no podríamos agruparlos por zonas. Ahora, la estructura de la base de datos se puede observar en la Figura 3.

	position	score	ratings	price_range	zip_code	lat	lng
0	19	4.7	200.0	$	35207.0	33.6	-86.8
1	9	4.7	200.0	$	35217.0	33.6	-86.8
2	6	4.7	200.0	$	35204.0	33.5	-86.9
3	17	4.7	200.0	$	35226.0	33.4	-86.8
4	162	4.7	22.0	$	35203.0	33.5	-86.8

Figura 3: Limpieza de la base de datos.

3.7 Tipos de variables

- Variables categóricas: price_range.
- Variables numéricas: position, score, ratings, zip_code, lat, lng.

4 Análisis exploratorio de datos

4.1 Gráfico de barras

Se realizó un gráfico de barras de la variable categórica 'price_range', el cual se observa a continuación en la Figura 4.

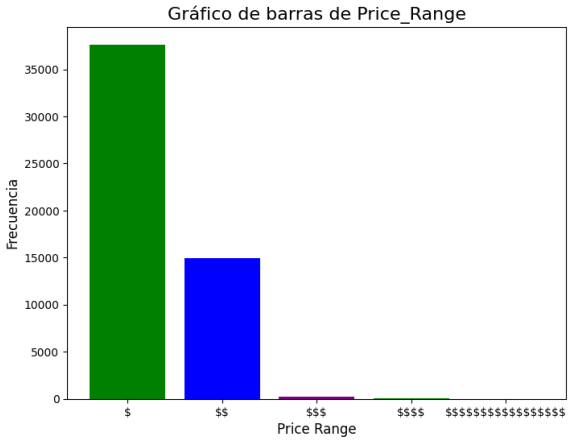

Figura 4: Se observa que hay una mayor frecuencia en la categoría de $ en comparación con las demás categorías.

4.2 Histogramas

En las siguientes figuras se observan los histogramas de las variables numéricas.

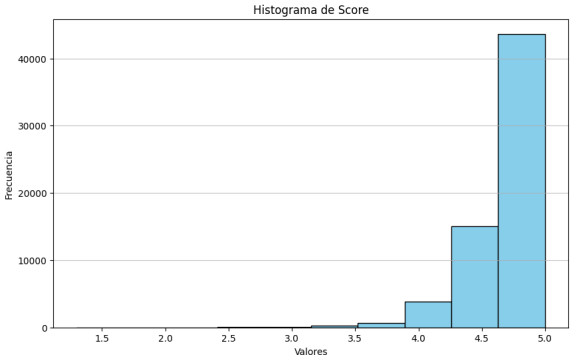

Figura 5: Los valores van de entre 1.5 a 5, con una frecuencia mayor entre 4.5 y 5.0.

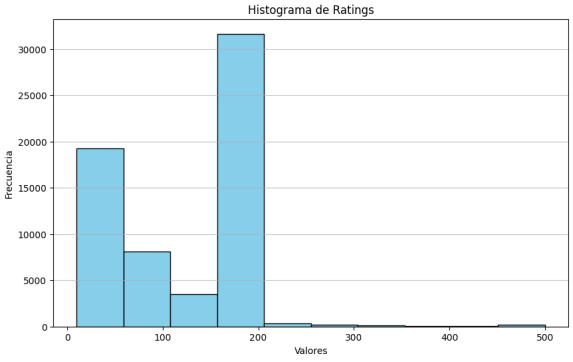

Figura 6: Los valores van de 0 a 500, con una mayor frecuencia entre 150 y 200.

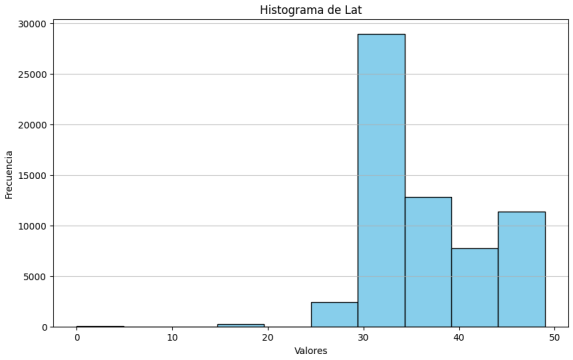

Figura 7: Los valores van de 0 a 50, con una mayor frecuencia entre 30 y 35.

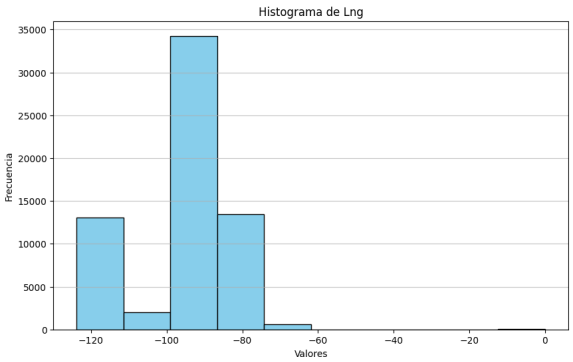

Figura 8: Distribución de longitudes: valores de -120 a 0, con mayor frecuencia entre -100 y -90.

4.3 Boxplots

Se presenta la Figura 9, que muestra los box plots correspondientes a las variables numéricas del conjunto de datos. En ellos se puede observar la distribución de las siguientes variables:

Score: La mediana de la variable es 4.6, con una baja variabilidad, lo que sugiere que las calificaciones son relativamente uniformes entre los restaurantes. Ratings: La mediana de esta variable es de 200, pero presenta una alta variabilidad, lo que indica que algunos restaurantes tienen significativamente más reseñas que otros. Lat: Los restaurantes están mayormente concentrados en una latitud entre 30 y 40, lo que sugiere que la mayoría se encuentran en zonas geográficas cercanas. Lng: En cuanto a la longitud, la mayor concentración se encuentra entre -100 y -90, lo que refleja una distribución regional de los restaurantes en ciertas áreas geográficas. Este análisis gráfico es útil para observar la dispersión y la presencia de valores atípicos en las variables numéricas.

4.4 Correlación y Relaciones entre Variables

- Mapa de calor de correlación: Se obtuvo el mapa de calor de correlación (Figura 10) entre las variables numéricas.

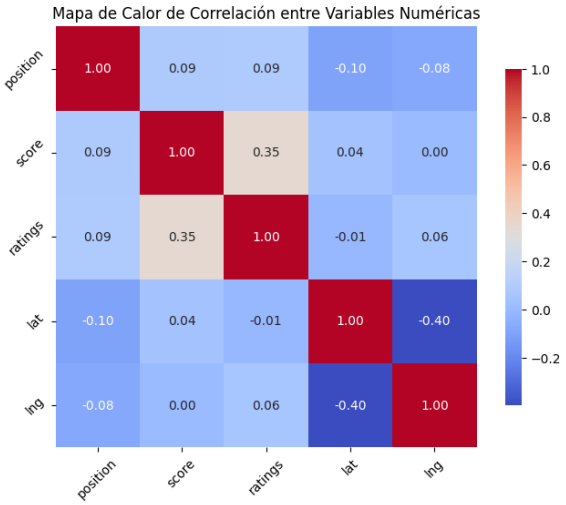

Figura 10: La mayoría de las correlaciones entre las variables son muy débiles. Sin embargo, se puede destacar la correlación positiva entre ratings y score, lo que indica que los restaurantes con un mayor número de calificaciones tienden a obtener mejores puntuaciones. Esto podría sugerir que los usuarios tienden a dejar calificaciones más altas en restaurantes que han recibido una mayor cantidad de atención o interacción. En cuanto a la correlación negativa entre latitud y longitud, esta relación podría explicarse por el hecho de que los restaurantes fueron seleccionados dentro de una área geográfica específica. Esto implica que la distribución de restaurantes podría estar limitada a ciertas zonas, lo que produce una relación inversa entre estas dos variables geográficas.

- Diagramas de dispersión Se realizaron los siguientes diagramas de dispersión, que se presentan en la Figura 11, para observar la relación entre las variables numéricas que tuvieron

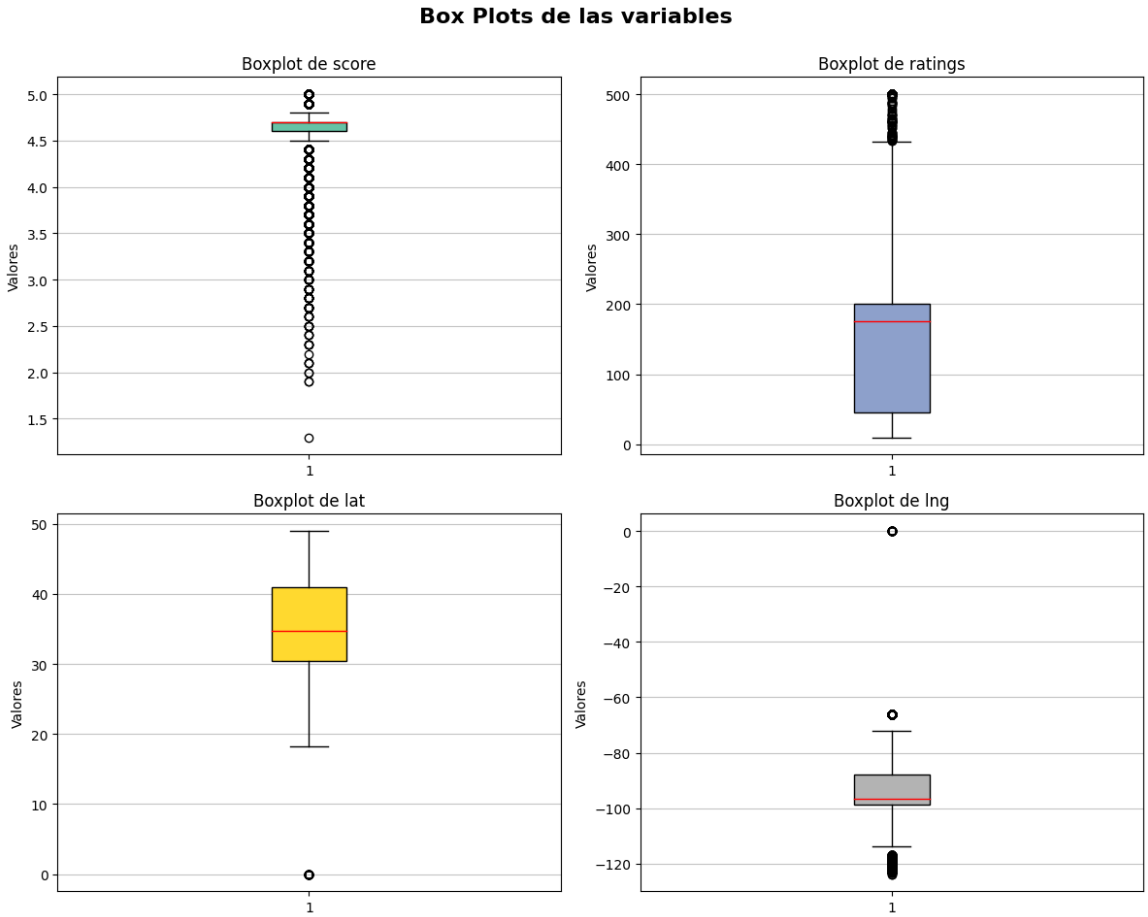

Figura 9: Los box plots muestran la distribución de las variables numéricas: **Score:** Mediana de 4.6, con poca variabilidad, lo que refleja uniformidad en calificaciones. **Ratings:** Mediana de 200, pero con alta variabilidad entre restaurantes. **Lat:** Restaurantes concentrados entre 30 y 40 de latitud. **Lng:** Mayor concentración entre -100 y -90 de longitud, reflejando una distribución regional.

una correlación considerable.

4.5 Descubrimiento de insights

Del análisis exploratorio de datos (EDA), se pudo observar que los restaurantes con una calificación alta, cerca de 4.6, tienden a tener una gran cantidad de calificaciones. Esto podría interpretarse como una indicación de que estos restaurantes son populares y reciben una valoración positiva constante por parte de los usuarios. Es probable que los usuarios de la plataforma confíen en estos restaurantes, lo que genera un ciclo donde las buenas calificaciones atraen más usuarios, lo que a su vez aumenta aún más las calificaciones.

Por otro lado, se observó que los restaurantes en la base de datos están concentrados en ciertas zonas geográficas específicas. Sin embargo, esto no implica necesariamente que los restaurantes en estas áreas sean los mejores. Es posible que la base de datos haya sido creada seleccionando restaurantes ubicados en zonas particulares, lo que podría haber influido en la distribución geográfica de los datos.

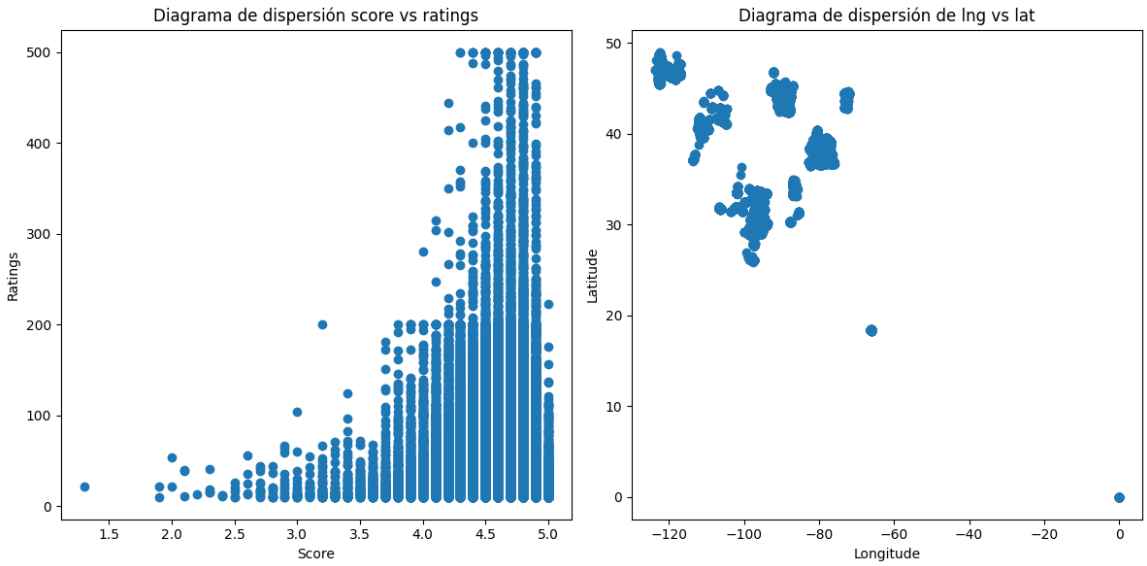

Figura 11: El gráfico de dispersión entre score y ratings muestra que, a medida que aumenta el score, también lo hacen las calificaciones, aunque con gran dispersión. Por otro lado, el gráfico de lng y lat revela que los restaurantes están concentrados en áreas geográficas específicas, sugiriendo que la ubicación influye en la distribución de las calificaciones.

Esta concentración geográfica podría ser el resultado de un sesgo en la recolección de los datos y no una verdadera representación de la calidad de los restaurantes en todo el país.

Además, durante el análisis se identificaron outliers en las variables Score, Ratings y lng, lo cual podría afectar la interpretación de los datos y los modelos predictivos. Estos outliers pueden representar casos atípicos que no siguen las mismas tendencias que el resto de los datos, lo que puede influir en la precisión de cualquier modelo que se construya si no se gestionan adecuadamente.

5 Ingeniería de características

- Se eliminaron los outliers de las variables *zona*, *score* y *lng* con la finalidad de evitar que afectaran la precisión de los modelos ni aumentaran la variabilidad de los datos. El procedimiento para eliminar los outliers consistió en identificar aquellos valores que quedaban fuera de los límites establecidos por los box plots, es decir, aquellos valores debajo de $Q1 - 1{,}5 \times \text{IQR}$ y aquellos por encima de $Q3 + 1{,}5 \times \text{IQR}$.

- Se realizó una estandarización de las columnas *ratings*, *lat* y *lng* para asegurar que estas variables estuvieran en el mismo rango y no distorsionaran los resultados de los modelos posteriores. Esto fue crucial para evitar que una variable, como *lng*, con valores más grandes, dominara sobre las demás, como *lat*, cuando se utilizaran en el análisis o en modelos predictivos.

- Se realizó una reducción de dimensionalidad utilizando PCA (Análisis de Componentes Principales) sobre las variables *lat* y *lng*, con el objetivo de crear una nueva variable llamada *zona* que representara la localización

geográfica de los restaurantes. La estandarización previa de estas variables fue fundamental para evitar que los valores de *lng* tuvieran un impacto desproporcionado en los resultados del PCA.

- Debido al desbalance en los datos, ya que había muchos más restaurantes considerados populares que no populares, se realizó un submuestreo de los restaurantes no populares. Esto ayudó a equilibrar las clases y evitar que los modelos se inclinaran demasiado hacia la clase mayoritaria, lo que podría haber afectado la precisión de las predicciones.

- A partir de la variable *score* y el insight de que un restaurante con un *score* superior a 4.6 se considera popular, se creó una nueva variable binaria llamada *popular*. A los restaurantes con un *score* inferior a 4.6 se les asignó un valor de 0, y a los que tenían un *score* superior a 4.6 se les asignó un valor de 1. Esta variable binaria se utilizaría en modelos de clasificación para predecir la popularidad de los restaurantes.

- Se realizó una codificación nominal de la variable *price_range* para poder trabajar con valores numéricos en modelos futuros. Esta transformación permite que las categorías de *price_range* se puedan utilizar de manera eficiente en modelos que requieren datos numéricos.

Con todo esto, el DataFrame se puede visualizar en la Figura 12.

	position	score	ratings	price_range_ordinal	zona	popular
0	-0.881505	4.7	0.680698	1.0	0.204962	1
1	-1.006841	4.7	0.680698	1.0	0.204962	1
2	-1.044442	4.7	0.680698	1.0	0.183072	1
3	-0.906572	4.7	0.680698	1.0	0.175508	1
4	0.910806	4.7	-1.788303	1.0	0.190235	1

Figura 12: DataFrame después de la ingeniería de características.

5.1 Justificación del Target y Selección de Variables

El target, definido como la variable binaria popular (1 para calificaciones ¿4.6 y 0 en caso contrario), se construyó con base en el EDA, donde se observó que los restaurantes con altas calificaciones y numerosas reseñas son percibidos como confiables y populares. Este enfoque simplifica el problema y alinea con el objetivo de proporcionar recomendaciones claras a los usuarios.

Las variables predictoras seleccionadas son:

- ratings: Indica el número de reseñas, correlacionado con la popularidad.

- price range: Representa el segmento económico del restaurante, relevante para las preferencias de los usuarios.

- zona: Combina ubicación geográfica (latitud y longitud), un factor clave en la elección del cliente.

6 Modelado de Datos

A continuación se describen los modelos utilizados en este proyecto, destacando sus características y aplicabilidad en el contexto de la clasificación de restaurantes.

- **Regresión logística**

La regresión logística es un modelo estadístico que permite clasificar una variable objetivo en función de una o más variables predictoras. Este modelo utiliza una función logística (sigmoide) para realizar las clasificaciones, asignando el valor 0 cuando la probabilidad de pertenecer a la clase es menor a 0.5 y el valor 1 cuando la probabilidad es mayor a 0.5. Este modelo es muy popular debido a su implementación sencilla y su capacidad para proporcionar una interpretación clara de los resultados.

La regresión logística es particularmente útil para tareas de clasificación binaria, como es el caso de este proyecto, donde se clasifican los restaurantes como *populares* (valor 1) o *no populares* (valor 0), en función de sus calificaciones y número de reseñas.

- **Árboles de decisión**

  Los árboles de decisión son modelos de clasificación que realizan particiones recursivas de los datos mediante pruebas en las características de las variables predictoras, generando una estructura jerárquica en forma de árbol. Cada nodo del árbol representa una decisión basada en una característica, y las ramas conducen a un resultado final (la clase a la que pertenece el restaurante).

  Este modelo es adecuado tanto para variables numéricas como categóricas. En este caso, se emplean las variables numéricas *popularidad*, *zona* y la variable categórica *price_range_ordinal* para generar las particiones y clasificar los restaurantes.

- **Máquinas de Vector de Soporte (SVM)**

  El modelo de Máquinas de Vector de Soporte busca encontrar un hiperplano que separa las diferentes clases en el espacio de características. Este hiperplano maximiza el margen entre las clases, y las distancias más cercanas entre los puntos de datos y el plano se denominan *vectores de soporte*.

  Las SVM son particularmente útiles cuando las relaciones entre las variables no son lineales, lo que es común en muchos problemas reales. En este caso, las variables de entrada no presentan correlaciones lineales entre sí, lo que hace que las SVM sean una opción adecuada para manejar la complejidad de los datos.

6.1 Propuesta de solución

Dado que la base de datos disponible no contiene información específica sobre los clientes, como sus preferencias o historial de compras, no es posible construir un modelo personalizado que recomiende restaurantes de acuerdo con los gustos individuales de cada usuario. Sin embargo, una solución efectiva sería clasificar los restaurantes en dos categorías: *populares* y *no populares*. Esta clasificación puede proporcionar a los consumidores una referencia valiosa a la hora de elegir un restaurante, especialmente para aquellos que buscan opciones con altas calificaciones o con una gran cantidad de reseñas.

A través de los modelos mencionados previamente, como la regresión logística, los árboles de decisión y las máquinas de soporte vectorial, se puede predecir esta clasificación de manera efectiva. Los modelos entrenados utilizando las características de los restaurantes, como las calificaciones, el número de reseñas y su ubicación, pueden generar una clasificación

confiable, ofreciendo a los consumidores una herramienta útil para la toma de decisiones.

Aunque la falta de datos de clientes limita las recomendaciones personalizadas, la clasificación de restaurantes en *populares* y *no populares* sigue siendo un paso significativo hacia la mejora de la experiencia del usuario, ya que les permite encontrar restaurantes de alta calidad de manera más eficiente. Los modelos desarrollados hasta el momento, junto con los análisis realizados en el proceso, son fundamentales para lograr esta clasificación y proporcionar un valor agregado a los usuarios de la aplicación.

7 Resultados y evaluación

En el Cuadro 1 se presentan los coeficientes del modelo de Regresión Logística, los cuales indican la relación entre cada variable predictora y la probabilidad de que un restaurante sea clasificado como *popular*. Estos coeficientes son fundamentales para interpretar cómo las variables, como la *popularidad*, *zona* y *price_range_ordinal*, afectan la clasificación. Los coeficientes positivos indican una mayor probabilidad de ser clasificado como *popular*, mientras que los coeficientes negativos sugieren lo contrario.

En el Cuadro 2 se presentan los coeficientes de importancia de las características en los Árboles de Decisión. Estos coeficientes reflejan el impacto de cada característica en la toma de decisiones dentro del modelo. Las características con valores más altos de importancia indican que tienen un mayor efecto en la predicción del modelo. La interpretación de estos coeficientes es clave para entender qué factores son los más determinantes en la clasificación de los restaurantes como populares o no populares.

Cabe destacar que no se pudieron obtener los coeficientes del modelo de Máquinas de Vector de Soporte (SVM) debido a que se utilizó un kernel no lineal. Este tipo de kernel mapea los datos a un espacio de características de mayor dimensión, lo que dificulta la interpretación directa de los coeficientes, ya que el modelo no produce una representación explícita en el espacio original de las características.

En el Cuadro 3 se presentan las métricas de rendimiento de los modelos, donde se observa que la exactitud de los tres modelos es bastante similar, con la Regresión Logística ligeramente superando a los demás. Los tres modelos logran un rendimiento de predicción correcto cercano al 80 %. Esto indica que los modelos son consistentes al hacer predicciones correctas, aunque con una ligera ventaja para la Regresión Logística.

En cuanto a la precisión, todos los modelos muestran un desempeño excelente, superando el 90 %. Sin embargo, el modelo SVM se destaca ligeramente al obtener la mayor precisión, lo que sugiere que es más capaz de distinguir correctamente entre las clases de restaurantes populares y no populares.

Además, los tres modelos presentan un buen desempeño en la identificación de instancias positivas (es decir, restaurantes populares), con tasas cercanas al 80 %. Entre ellos, la Regresión Logística y los Árboles de Decisión tienen un rendimiento superior, lo que significa que ambos son mejores para identificar correctamente los restaurantes populares.

En el Cuadro 4 se observa que los tres modelos tienen un buen rendimiento global, con promedios cercanos a 0.9 en las métricas de evaluación, como la precisión y la puntuación F1, lo que indica que los modelos están generalizando bien. Además, muestran una baja variabilidad,

lo que sugiere que los resultados son consistentes entre diferentes subconjuntos de datos, lo que es un buen indicio de que los modelos no están sobreajustados.

Finalmente, en la Figura 13 se muestra la comparación de los tres modelos utilizando la curva ROC, que es una herramienta útil para evaluar la capacidad de clasificación binaria. Los valores de área bajo la curva (AUC) se acercan a 0.8 para todos los modelos, con el SVM alcanzando el valor más alto de 0.86. Esto refuerza la idea de que todos los modelos tienen un buen desempeño en términos de clasificación, pero el SVM es ligeramente superior, lo que lo convierte en el mejor modelo para este caso específico.

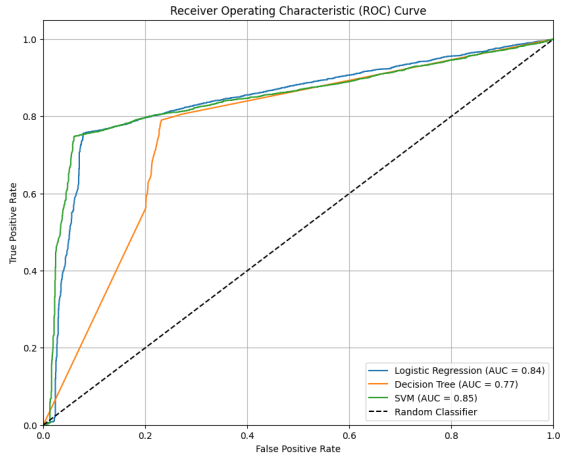

Figura 13: ROC curve de los modelos

8 Conclusión

En este proyecto se desarrollaron y evaluaron modelos de predicción para clasificar a los restaurantes como *populares* o *no populares*, basándose en varias características clave como la zona geográfica donde se localizan, el número de reseñas que han recibido y el rango de precio de los restaurantes. Estos factores fueron seleccionados por su posible influencia en la percepción de los clientes sobre la popularidad de un restaurante, y se utilizaron en combinación para entrenar tres modelos principales: Regresión Logística, Árboles de Decisión y Máquinas de Vector de Soporte (SVM).

Los resultados obtenidos muestran que todos los modelos tuvieron un rendimiento muy similar, con la Regresión Logística mostrando una ligera ventaja en cuanto a exactitud y la precisión del modelo SVM siendo marginalmente superior. En general, los modelos lograron identificar correctamente a los restaurantes populares con una precisión mayor al 80 % y un buen rendimiento en términos de métricas como precisión y puntuación F1. Estos hallazgos indican que las variables seleccionadas, como la zona, las reseñas y el rango de precio, son buenas indicadores de la popularidad de los restaurantes y que los modelos son adecuados para clasificar con eficacia estos datos.

La clasificación de los restaurantes en *populares* y *no populares* puede ser extremadamente útil para la aplicación Uber Eats. Al incorporar este sistema de recomendaciones, Uber Eats podría ofrecer a sus usuarios sugerencias más acertadas basadas en la popularidad de los restaurantes, mejorando la experiencia del cliente. Este tipo de clasificación también tiene el potencial de asegurar una compra, ya que los usuarios tienden a sentirse más atraídos por restaurantes bien valorados y populares. Además, al incentivar la compra de estos restaurantes populares, Uber Eats podría incrementar sus ventas y fortalecer su posición en el mercado de servicios de entrega de alimentos, lo que podría resultar en una mayor retención de usuarios y un aumento en la satisfacción general del cliente.

En resumen, la capacidad de predecir si un restaurante será popular o no utilizan-

Variable	Coeficiente de Regresión Logística
Variable 1: 'zona'	0.11
Variable 2: 'ratings'	1.40
Variable 3: 'price_range_ordinal'	0.01

Cuadro 1: Coeficientes de la regresión logística.

Variable	Importancia en el Árbol de Decisión
Variable 1: 'zona'	0.24
Variable 2: 'ratings	0.73
Variable 3: 'price_range_ordinal'	0.02

Cuadro 2: Importancia de las variables en el Árbol de Decisión.

Métrica	Regresión Logística	Árbol de Decisión	SVM
Exactitud	0.79	0.78	0.78
Precisión	0.94	0.93	0.96
Recall	0.79	0.79	0.76

Cuadro 3: Comparación de Modelos

Parámetro	Regresión logística	Árboles de decisión	SVM
Precisión Promedio	0.86	0.88	0.87
Desviación Estándar	0.02	0.03	0.01

Cuadro 4: Resultados de la validación cruzada para los modelos.

do características clave proporciona una ventaja estratégica para Uber Eats al mejorar la calidad de las recomendaciones a sus clientes, asegurando una experiencia de usuario más personalizada y eficiente, y contribuyendo al crecimiento continuo de la plataforma.

Repositorio del proyecto

El siguiente código QR lleva directo al respositorio de este proyecto en GitHub.

Figura 14: Escanéando este código QR lo llevará al repositorio del proyecto.

Referencias

1. Statista. (s.f.). Online Food Delivery - Worldwide. Recuperado 14 de septiembre de 2024, de https://www.statista.com/outlook/emo/online-food-delivery/worldwide

2. Grand View Research (s.F) Online Food Delivery Services Market Size, Share & Trends Analysis Report By Channel Type (Mobile Application, Websites/Desktop), By Payment Method (COD, Online), By Type, By Region, And Segment Forecasts, 2022 - 2030. Recuperado el 14 de septiembre de https://www.grandviewresearch.com/industry-analysis/online-food-delivery-services-market

3. Far Eye (s.f) On-demand food delivery software to resolve food delivery challenges. Recuperado el 14 de septiembre de 2024 de https://fareye.com/resources/blogs/on-demand-food-delivery-industry

4. Altametrics (s.f) Top Challenges in Food Services Delivery and How to Overcome Them. Recuperado el 14 de septiembre de 2024 de https://altametrics.com/topics/top-challenges-in-food-services-delivery-and-how-to-overcome-them/

5. Batra, B. (2023, 26 febrero). 9 Key Benefits of Online Ordering for Restaurants. Resto Labs. Recuperado el 14 de septiembre de 2024, de https://www.restolabs.com/blog/advantages-food-ordering-system-restaurants

6. Restolabs Team (S.f). How data analytics can help restaurants generate higher revenue. Recuperado el 14 de septiembre de 2024, de https://www.restolabs.com/blog/how-data-analytics-can-help-restaurants-generate-higher-revenue

7. McMahon D. (2022, August). Top Challenges in Food Services Delivery and How to Overcome Them. Recuperado el 14 de septiembre de 2024, de https://altametrics.com/topics/top-challenges-in-food-services-delivery-and-how-to-overcome-them/

8. 42 Works (2024, July 25). Top Six Challenges Faced By Online Food Delivery Services. Recuperado el 14 de septiembre de 2024, de https://42works.net/top-six-challenges-faced-by-online-food-delivery-services/

9. Curry, D. (2024, October 30). Uber Eats Revenue and Usage Statistics (2024). Recuperado el 14 de septiembre de 2024, de https://www.businessofapps.com/data/uber-eats-statistics/

10. Mulla, R. [Rob Mulla]. (2021, 31 diciembre). Exploratory Data Analysis with Pandas Python [Vídeo]. YouTube. Recuperado 23 de octubre de 2024, de https://www.youtube.com/watch?v=xi0vhXFPegw&t=1716s

11. Edureka! (2020, Abril 29). Exploratory Data Analysis (EDA) Using Python — Python Data Analysis — Python Training — Edureka [Vídeo]. YouTube. Recuperado el 23 de octubre de 2024, de https://www.youtube.com/watch?v=-o3AxdVcUtQ

12. Sotaquirá, M. [Codificando Bits]. (2023, Mayo 26). Algoritmos y Modelos de Machine Learning [Vídeo].

YouTube. Recuperado el 29 de noviembre de 2024, de https://www.youtube.com/watch?v=tGU1vzdjMfI&t=360s

13. González L. [AprendeIA con Ligdi Gonzalez]. (2019, Mayo 24). MÉTRICAS DE EVALUACIÓN MODELOS DE CLASIFICACIÓN SCIKIT LEARN — #35 Curso Machine Learning con Python [Vídeo]. YouTube. Recuperado el 29 de noviembre de 2024, de https://www.youtube.com/watch?v=K5PNrX694HQ&list=PLJjOveEiVE4Dk48EI7I-67PEleEC5nxc3&index=36