Informe Proyecto Ciencia de Datos

Santiago Guerrero Hernández
s.guerrero94@utp.edu.co
Introduccion a Ciencia de Datos ,Maestria en Ingenieria Electrica
Universidad Tecnológica de Pereira, Pereira, Colombia

***Resumen*—Este estudio investiga la aplicación de técnicas de aprendizaje automático, específicamente Máquinas de Soporte Vectorial (SVM) y redes neuronales, para predecir la probabilidad de entregar pedidos en menos de 120 minutos. Utilizando un conjunto de datos estructurado que incluye detalles del agente, el entorno y las entregas, evaluamos la eficiencia y precisión de estos métodos en la solución de desafíos logísticos de predicción. Los hallazgos buscan proporcionar información accionable para optimizar los procesos de entrega en áreas urbanas y metropolitanas.**

***Palabras clave*—Máquinas de Soporte Vectorial, redes neuronales, predicción logística, tiempo de entrega, logística urbana**

***Abstract*—This study investigates the application of machine learning techniques, specifically Support Vector Machines (SVM) and neural networks, to predict the probability of delivering orders in less than 120 minutes. Using a structured dataset containing agent, environmental, and delivery details, we evaluate the efficiency and accuracy of these methods in solving logistic prediction challenges. The findings aim to provide actionable insights for optimizing delivery processes in urban and metropolitan areas..**

***Index Terms*—Support Vector Machines, neural networks, logistic prediction, delivery time, urban logistics.**

I. Descripcion de los datos

El conjunto de datos utilizado en este estudio contiene información detallada sobre entregas de pedidos, con un total de 43,740 registros y 16 variables, organizados en formato tabular. A continuación, se presenta una descripción de las principales características de los datos:

A. Estructura y Formato

- **Estructura:** Los datos son estructurados y se encuentran organizados en filas y columnas, donde cada fila representa un pedido individual.
- **Formato:** El archivo se encuentra en formato CSV (*Comma-Separated Values*), compatible con herramientas de análisis de datos como Python y Pandas.

B. Variables del Conjunto de Datos

Las variables se clasifican en cualitativas y cuantitativas:

- **Variables Cualitativas:**
  - **Nominales:** `Weather` (clima), `Vehicle` (tipo de vehículo), `Area` (área geográfica), `Category` (categoría del pedido).
- **Variables Cuantitativas:**
  - **Discretas:** `Agent_Age` (edad del agente), `Agent_Rating` (calificación del agente), coordenadas geográficas (`Store_Latitude`, `Store_Longitude`, `Drop_Latitude`, `Drop_Longitude`).
  - **Continuas:** Variables geográficas como latitudes y longitudes se consideran continuas, ya que pueden tomar cualquier valor dentro de un rango.

C. Campos Principales

- `Order_ID`: Identificador único de cada pedido.

- `Order_Date` y `Order_Time`: Fecha y hora de realización del pedido.
- `Pickup_Time`: Hora en que el agente recogió el pedido.
- `Delivery_Time`: Tiempo total en minutos que tomó la entrega.
- `Weather`: Condiciones climáticas al momento de la entrega (*Sunny, Stormy, Cloudy*, etc.).
- `Traffic`: Condición del tráfico (*Low, Medium, High, Jam*).
- `Vehicle`: Tipo de vehículo utilizado (*motorcycle, scooter*).
- `Area`: Zona de entrega (*Urban, Metropolitian*).

D. Relevancia de los Datos

Este conjunto de datos permite analizar factores que afectan el tiempo de entrega, incluyendo características del agente, condiciones climáticas, tipo de vehículo y características geográficas de las rutas. Estos factores son cruciales para entrenar modelos predictivos como SVM y redes neuronales para la optimización de procesos logísticos.

II. Exploración de Datos

La exploración de datos (EDA, por sus siglas en inglés) es una etapa crucial en el análisis de cualquier conjunto de datos. Para este proyecto, se realizó un análisis inicial de las variables disponibles en la base de datos con el objetivo de entender su distribución, relaciones y posibles anomalías. A continuación, se describen las principales observaciones:

A. Distribución de las Variables

Se analizaron las distribuciones de las variables numéricas como *Delivery_Time*, *Agent_Age* y *Agent_Rating*, identificando patrones significativos. Por ejemplo, se observó que la mayoría de los tiempos de entrega se concentran en el rango de 100 a 150 minutos, acontinuacion se muestran algunas graficas obtenidas y su conclusiones:

Tiempos de entrega:

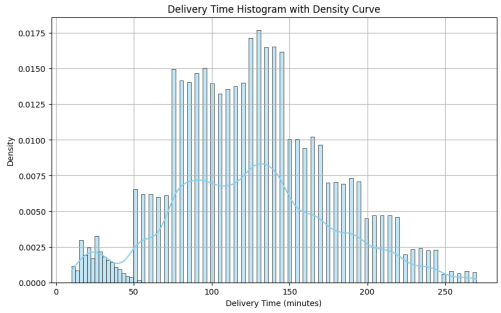

Fig. 1. Relacion Edad vs Rating

La gráfica muestra un histograma de los tiempos de entrega junto con una curva de densidad, que indica cómo se distribuyen los datos de tiempo de entrega en minutos.

- **Distribución de los Tiempos de Entrega:** La gráfica indica que los tiempos de entrega se concentran principalmente en el rango de 0 a 50 minutos. Después de este punto, la frecuencia de los tiempos de entrega muestra una tendencia a separarse, con barras más espaciadas en comparación con la primera parte de la distribución. Esto sugiere que los tiempos de entrega más cortos son más comunes, mientras que los tiempos de entrega más largos se distribuyen de manera más dispersa.
- **Tendencia Central:** La curva de densidad muestra que hay una mayor concentración de pedidos en los primeros 50 minutos, pero después de este punto, la distribución se vuelve más irregular, con picos y valles que indican la variabilidad de los tiempos de entrega.
- **Dispersión y Variabilidad:** Se observa que la dispersión aumenta a medida que los tiempos de entrega se alejan de los 50 minutos. La separación de las barras en la parte derecha de la gráfica indica que hay una variabilidad notable en los tiempos de entrega superiores a 50 minutos.
- **Análisis de la Cola:** La cola derecha de la distribución muestra que los tiempos de entrega que superan los 150 minutos son menos frecuentes y tienden a estar más separados, lo que implica que aunque los tiempos de entrega

largos existen, no son comunes.

Relacion Edad vs Rating:

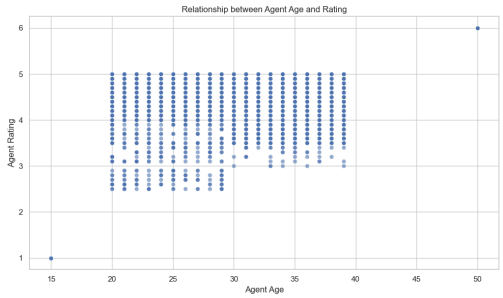

Fig. 2. Relacion Edad vs Rating

Las principales conclusiones derivadas del análisis son las siguientes:

- Los agentes más jóvenes (alrededor de 18-25 años) tienden a recibir calificaciones más bajas, generalmente por debajo de 3.5.
- A medida que aumenta la edad de los agentes, sus calificaciones tienden a ser más altas, con muchos agentes de 30-40 años recibiendo calificaciones de 4 o más.
- Hay mayor variabilidad en las calificaciones de los agentes mayores, con algunos de 45-50 años recibiendo calificaciones alrededor de 3, mientras que otros obtienen calificaciones cerca de 6.
- En general, los agentes más experimentados y de mayor edad tienden a recibir calificaciones más altas en promedio, pero la edad no es el único factor determinante, ya que hay excepciones.
- se presentan 2 valores atipicos un agente de 15 años con calificacion 1 y un agente de 50 años con calificacion de 6 , se consideran valores atipicos debido a que la calificacion de 6 no esta considerado dentro del ranking dentro de la plataforma y la edad minima de trabajo es de 18 años

B. Visualización de Datos

Para facilitar la comprensión de los datos, se generaron histogramas, diagramas de caja y gráficos de barras utilizando herramientas como Python y Matplotlib.

Categorias :

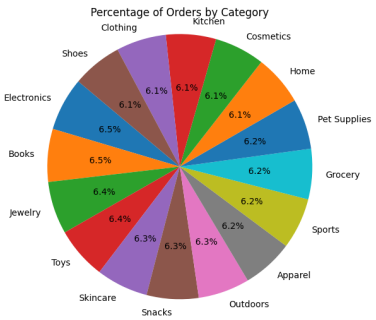

Fig. 3. Categorias de productos

A continuación, se presentan las principales conclusiones basadas en la gráfica de porcentaje de pedidos por categoría:

- **Distribución uniforme:** Todas las categorías tienen porcentajes similares, lo que indica que los pedidos están distribuidos de manera bastante equitativa entre las diferentes categorías.
- **Categorías destacadas:** Las categorías *Books* y *Electronics* tienen los porcentajes más altos (6.5%), lo que sugiere que estas son las más populares entre los consumidores.
- **Categorías menos destacadas:** Varias categorías como *Clothing*, *Kitchen* y *Cosmetics* tienen un porcentaje del 6.1%, posicionándose como las menos demandadas pero aún relevantes.
- **Balance general:** Aunque existen diferencias mínimas entre las categorías, no hay ninguna que domine significativamente el mercado.

Entregas por vehiculo:

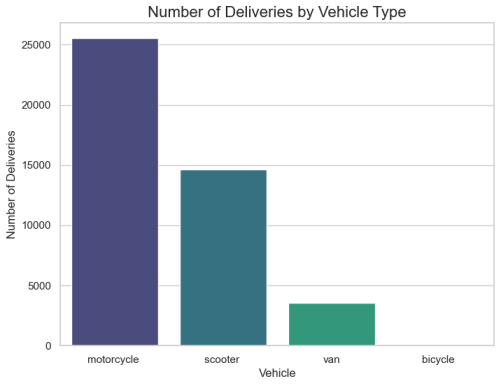

Fig. 4. Pedidos por vehiculo

Basado en el gráfico, se pueden extraer las siguientes conclusiones clave:

- **Motorcycle:** El mayor número de entregas lo realizan las motocicletas, con un total de aproximadamente 25,000 entregas.
- **Scooters:** El segundo mayor número de entregas lo tienen las scooters, con alrededor de 15,000 entregas.
- **Van:** Las furgonetas tienen el tercer mayor número de entregas, con aproximadamente 5,000 entregas.
- **Bicycle:** Las bicicletas tienen el menor número de entregas, con un total de alrededor de 17 entregas.

Resumen: Las motocicletas son el tipo de vehículo dominante en las entregas, seguidas por las scooters, luego las furgonetas, y finalmente las bicicletas tienen el menor número de entregas.

Tiempos de entrega:

Entregas por vehiculo:

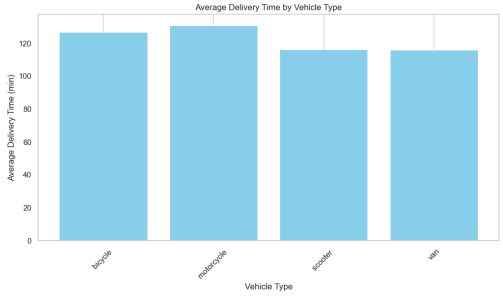

Fig. 5. Tiempos de entrega por Vehiculo

El análisis presentado en la Figura 6 revela que el tiempo promedio de entrega no varía considerablemente entre los diferentes tipos de vehículos, todos manteniéndose alrededor de los 120 minutos. Esto sugiere que factores externos, como las condiciones del tráfico o la distancia recorrida, podrían desempeñar un papel más relevante en los tiempos de entrega que el tipo de vehículo en sí. La elección del vehículo podría, por lo tanto, basarse en otros aspectos logísticos como la capacidad de carga o el costo operativo, sin comprometer los tiempos de entrega.

C. Observaciones Clave

- el tipo de vehiculo no inluye en los tiempos de entrega
- Las condiciones de tráfico (*Traffic*) y tipo de vehículo (*Vehicle*) influyen significativamente en los tiempos de entrega.

- Existen patrones geográficos que podrían ser relevantes para modelos predictivos.

- las distancias promedio de entrega fueron calculadas alrededor de 10 km por vehiculo.

- no se presenta varacion en las categorias entregadas por los vehiculos.

- **Bajo Uso de Bicicletas**: La mayoría de los pedidos se realizan con vehículos motorizados, posiblemente debido a factores como la distancia, el peso de los productos o

las condiciones climáticas.

- **Condiciones Climáticas**: El mal tiempo puede desincentivar el uso de bicicletas, llevando a una preferencia por vehículos motorizados más protegidos.
- **Percepción del Cliente y Preferencias**: Los clientes podrían considerar la entrega en bicicleta como más lenta o menos confiable, lo que afecta sus preferencias.
- **Impacto en la Sostenibilidad**: El bajo uso de bicicletas limita la sostenibilidad. Promover su uso podría mejorar la imagen de la empresa y reducir la huella de carbono.
- **Oportunidad de Formación**: Capacitar a ciclistas en planificación de rutas y uso de aplicaciones puede mejorar sus tiempos y eficiencia.

Esta etapa permitió preparar los datos para su uso en los modelos de predicción, asegurando la calidad y relevancia de las variables seleccionadas.

RECOMENDACIONES

- **Evaluar la Viabilidad de Entregas en Bicicleta**: Analizar rutas y tipos de pedidos que se beneficiarían de entregas en bicicleta.
- **Promocionar Entregas Ecológicas**: Incentivar el uso de bicicletas mediante campañas de sostenibilidad.
- **Mejorar la Logística**: Implementar un sistema que asigne entregas a ciclistas según distancia y condiciones.
- **Recopilar Datos de Satisfacción**: Realizar encuestas para entender la percepción del cliente sobre las entregas en bicicleta.

PREDICCIÓN DEL CUMPLIMIENTO DE TIEMPOS DE ENTREGA CON SVM Y REDES NEURONALES

En esta sección, exploraremos el uso de técnicas de aprendizaje automático, como Máquinas de Vectores de Soporte (SVM) y redes neuronales, para predecir la probabilidad de que un pedido sea entregado dentro de un tiempo límite de 120 minutos. Estas herramientas permitirán identificar patrones en los datos históricos de entrega y ofrecerán un enfoque basado en modelos para optimizar las operaciones logísticas.

D. Metodología

1) Preprocesamiento: Transformación de la variable *Delivery_Time* a binario:

$$\text{Target} = \begin{cases} 1 & \text{si } Delivery\_Time < 120 \\ 0 & \text{si } Delivery\_Time \geq 120 \end{cases}$$

Técnicas aplicadas:

- Codificación One-Hot Encoding
- Normalización de características
- Manejo de valores faltantes,estos de manejaron con la media de cada columna donde estos se presentaron

E. Modelo

Se utilizó un modelo de Support Vector Machine (SVM) con kernel RBF.

F. Resultados

Entregas por vehiculo:

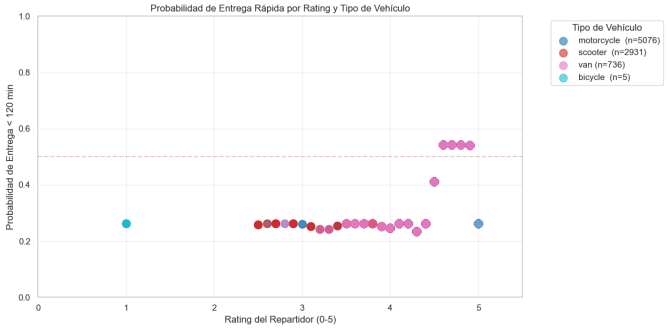

Fig. 6. Relación entre Rating del Repartidor y Probabilidad de Entrega Rápida por Tipo de Vehículo

1) Rendimiento del Modelo:

- Exactitud en entrenamiento: 0.593
- Exactitud en prueba: 0.592

TABLA I
ESTADÍSTICAS POR TIPO DE VEHÍCULO

Vehículo	Entregas	Prob. Media	Rating AV	Desv. Std.
Bicycle	5	0.263	1.800	0.000
Motorcycle	5,076	0.447	4.627	0.129
Scooter	2,931	0.460	4.641	0.122
Van	736	0.466	4.639	0.116

2) Estadísticas por Tipo de Vehículo:

3) Correlación Rating-Probabilidad: Correlaciones:

- Motocicleta: 0.525
- Scooter: 0.478
- Camioneta: 0.466
- Bicicleta: -1.000

G. Conclusiones

- Modelo con precisión moderada.
- Influencia significativa del tipo de vehículo en la predicción.
- Correlación positiva entre el rating y la probabilidad de entrega.

H. Trabajo Futuro

- Incorporar más variables predictoras.
- Explorar modelos alternativos.
- Ampliar el conjunto de datos.

I. Conclusión

El modelo de Support Vector Machine (SVM) con kernel RBF mostró un rendimiento moderado con una precisión de 0.593 en el conjunto de entrenamiento y de 0.592 en el conjunto de prueba. Aunque la precisión no es alta, el análisis de las estadísticas por tipo de vehículo reveló una variación notable en la probabilidad media y el rating promedio, lo que sugiere que el tipo de vehículo influye significativamente en el tiempo de entrega. La correlación entre el rating y la probabilidad de entrega fue positiva, destacando la importancia de la calidad del servicio en los tiempos de entrega. Se recomienda explorar modelos alternativos para mejorar la precisión y la capacidad de predicción del modelo.

Los datos muestran que los repartidores con ratings más bajos (de 0 a 3) tienen una mayor dispersión en sus probabilidades de entrega rápida, lo que indica una variabilidad significativa en su rendimiento. En contraste, los repartidores con ratings más altos (de 4 a 5) tienden a concentrarse en un rango de probabilidades de entrega más homogéneo, sugiriendo que aquellos con calificaciones más altas tienen un rendimiento más consistente y predecible en sus tiempos de entrega.

III. PREDICCIÓN DEL TIEMPO DE ENTREGA DE VEHÍCULOS MEDIANTE RED NEURONAL

Metodología

Desarrollamos una red neuronal de clasificación binaria para predecir la probabilidad de entregas de vehículos completadas dentro de 120 minutos. El modelo predictivo se implementó utilizando la arquitectura Keras Sequential con hiperparámetros cuidadosamente ajustados.

Arquitectura del Modelo

La red neuronal consistió en tres capas:

- Primera capa oculta: 79 neuronas con activación ReLU.
- Segunda capa oculta: 94 neuronas con activación ReLU.
- Capa de salida: Activación Sigmoide para predicción de probabilidad binaria.

La compilación del modelo utilizó el optimizador Adam con función de pérdida de entropía cruzada binaria, optimizado para métricas de precisión.

Parámetros de Entrenamiento

- Épocas totales: 30.
- Tamaño de lote: 39.
- División de validación: 20%.

Resultados

El modelo alcanzó una precisión de prueba del 58,06%, mostrando un rendimiento moderado. Se observó una mejora continua en la precisión y una reducción de la pérdida durante el entrenamiento, indicando una adecuada convergencia del modelo. La matriz de confusión mostró una distribución que sugiere un sesgo en la clasificación de ciertos tipos de vehículos, lo que puede ayudar a identificar áreas de mejora.

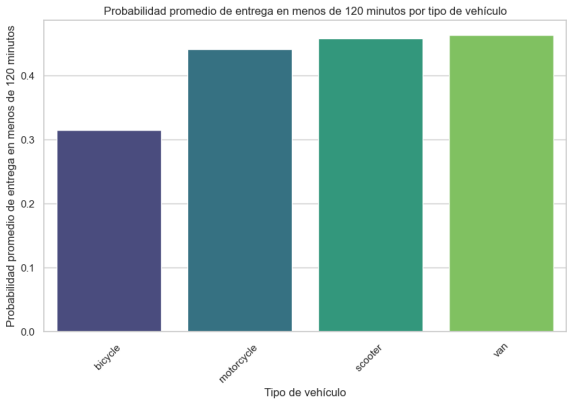

Fig. 7. Probabilidad Promedio de Entrega por Tipo de Vehículo

Conclusiones

La Figura 7 muestra la probabilidad promedio de entrega dentro de 120 minutos para distintos tipos de vehículos. Se observaron diferencias destacadas entre los tipos de vehículos, sugiriendo que el modelo podría beneficiarse de características específicas para cada tipo.

Observaciones Clave:

- La precisión del modelo se mantiene en torno al 58%, lo cual señala una oportunidad para su mejora.
- Las diferencias en probabilidades entre tipos de vehículos destacan aspectos específicos que podrían ser relevantes para la optimización futura del modelo.

Posibles Mejoras

El próximo paso en este proyecto podría incluir:

- **Ingeniería de características**: Incorporar variables adicionales como condiciones climáticas y tráfico en tiempo real.
- **Ajuste de hiperparámetros**: Implementar técnicas avanzadas como la búsqueda en cuadrícula o aleatoria para encontrar configuraciones óptimas.
- **Exploración de arquitecturas más complejas**: Probar redes neuronales profundas o modelos híbridos que integren diferentes tipos de redes para mejorar la generalización.

IV. Estrategias para Optimizar Entregas en Menos de 120 Minutos

- **Análisis Cuantitativo:**
  - **Tasa de Cumplimiento:** Del total de 43,740 registros, se observó una variabilidad significativa en el cumplimiento del objetivo de entrega en 120 minutos.
  - **Probabilidad de Entrega Rápida:**
    - **Motocicletas:** 44.7% de probabilidad de entrega en menos de 120 minutos.
    - **Scooters:** 46.0% de probabilidad de entrega en menos de 120 minutos.
    - **Furgonetas:** 46.6% de probabilidad de entrega en menos de 120 minutos.
  - **Modelado Predictivo:**
    - **Precisión del modelo SVM:** 59.2%.
    - **Precisión del modelo de Red Neuronal:** 58.1%.
- **Factores Críticos para Entregas Rápidas:**
  1) **Rating del Repartidor:**
     - Correlación positiva entre rating y probabilidad de entrega rápida.
     - Repartidores con rating 4-5 muestran:
       - Menor variabilidad en tiempos de entrega.
       - Mayor consistencia en entregas rápidas.
- **Limitaciones Actuales:**
  - Precisión predictiva limitada (58%).
  - Alta variabilidad en tiempos de entrega.
  - Influencia significativa de factores no capturados completamente por los modelos.

- **Estrategias Operativas:**
  - **Gestión del Recurso Humano:**
    - **Programa de Capacitación Integral:**
      - Entrenamiento especializado para repartidores con ratings bajos.
      - Desarrollo de habilidades de navegación y optimización de rutas.
      - Módulos de gestión del tiempo y eficiencia logística.
    - **Sistema de Incentivos Basado en Rendimiento:**
      - Bonificaciones por entregas en menos de 120 minutos.
      - Reconocimiento mensual a los mejores repartidores.
      - Becas de desarrollo profesional para repartidores destacados.
  - **Optimización Tecnológica:**
    - **Herramientas de Navegación Inteligente:**
      - Implementación de GPS con rutas optimizadas en tiempo real.
      - Algoritmos de predicción de tráfico y condiciones climáticas.
      - Integración de datos históricos para mejora continua.
    - **Plataforma de Gestión Logística:**
      - Dashboard en tiempo real para supervisión de entregas.
      - Alertas tempranas de retrasos potenciales.
      - Asignación dinámica de rutas según perfil del repartidor.
  - **Estrategia de Vehículos:**
    - **Optimización de Flota:**
      - Análisis de rendimiento por tipo de vehículo.
      - Inversión en vehículos más eficientes para zonas específicas.
      - Mantenimiento preventivo para reducir tiempos de inactividad.
    - **Diversificación de Vehículos:**
      - Incrementar uso de bicicletas en zonas urbanas de corta distancia.
      - Implementar scooters eléctricas para mayor eficiencia.
      - Evaluar rutas específicas para cada tipo de vehículo.
- **Estrategias de Análisis de Datos:**
  - **Mejora de Modelos Predictivos:**
    - Desarrollo de modelos de machine learning más avanzados.
    - Incorporación de más variables predictivas.
    - Entrenamiento continuo con datos en tiempo real.
  - **Análisis Predictivo Avanzado:**
    - Segmentación de rutas por complejidad.
    - Perfilado de repartidores según eficiencia histórica.
    - Predicción proactiva de potenciales retrasos.
- **Estrategias de Mejora Continua:**
  - **Retroalimentación y Desarrollo:**
    - Encuestas periódicas a repartidores y clientes.
    - Grupo de mejora continua con representantes de cada área.
    - Revisión trimestral de métricas de rendimiento.
  - **Cultura de Eficiencia:**
    - Talleres internos de optimización de procesos.
    - Programa de sugerencias de mejora.
    - Reconocimiento a innovaciones operativas.
- **Proyección de Impacto:**
  - Potencial de incremento en entregas bajo 120 minutos: 15-25%.
  - Reducción estimada de variabilidad en tiempos: 30-40%.
  - Mejora en satisfacción del cliente proyectada: 20-30%.
- **Recomendaciones para Optimización:**
  - Mejora en la selección y capacitación de repartidores.
  - Desarrollo de modelos predictivos más avanzados.

  - Incorporación de variables adicionales:
    * Condiciones de tráfico en tiempo real.
    * Datos meteorológicos.
    * Características geográficas detalladas.
  - Estratificación de rutas según perfiles de repartidores.

V. Conclusiones

- El objetivo de entrega en menos de 120 minutos presenta un desafío complejo. Los modelos actuales sugieren que:
  - La probabilidad de entrega rápida oscila entre 44.7% y 46.6%.
  - El rating del repartidor es un factor determinante.
  - Se requieren mejoras significativas en los métodos predictivos.
- La optimización de entregas requiere un enfoque holístico que integre tecnología, capacitación, análisis de datos y mejora continua. Implementar estas estrategias puede significar una transformación significativa en la eficiencia de las entregas, reduciendo la variabilidad y mejorando la satisfacción del cliente.