Caso de estudio: Accidentes de trafico

1. Objetivo del proyecto

*Desarrollar un **modelo predictivo de accidentes graves** que, a partir de los factores históricos y ambientales, estime la probabilidad de ocurrencia de accidentes con lesiones incapacitantes o fatales en cada tramo vial y franja horaria.*

Objetivos específicos:

1. *Analizar patrones históricos de accidentes graves según variables de vía, clima y hora.*
2. *Detectar variables que contribuyen más a la severidad de los accidentes.*
3. *Construir un modelo de clasificación que prediga la probabilidad de accidentes graves.*
4. *Proponer un **sistema de alertas y priorización de recursos** para la prevención de accidentes.*
5. *Validar la efectividad del modelo y generar métricas accionables (PR AUC, recall de accidentes graves).*

Preguntas de negocio:

Pregunta de negocio	***Métrica / Análisis Predictivo***	***Acción recomendada***
1. ¿Cuáles son los tramos viales y franjas horarias con mayor probabilidad de accidentes graves?	*Probabilidad de accidente grave por tramo vial × hora (modelo de clasificación, PR AUC, ranking de riesgo)*	*Priorizar patrullaje y campañas preventivas en tramos/hora de alto riesgo*
2. ¿Qué factores contribuyen más a la severidad de los accidentes*?*	*Feature importance / SHAP sobre modelo de severidad (variables como clima, iluminación, tipo de vía, defectos)*	*Implementar intervenciones focalizadas según causas críticas (p.ej. señalización, alertas climáticas)*
3. ¿Cómo afecta la combinación de condiciones climáticas, iluminación y tipo de vía al riesgo de accidentes graves?	*Análisis de interacción de variables en el modelo, heatmaps de riesgo*	*Diseñar políticas preventivas para condiciones específicas, como alertas de lluvia, visibilidad baja o curvas peligrosas*
4. ¿Qué impacto tendría la reparación de defectos viales o la mejora de señalización en la reducción de accidentes graves?	*Simulaciones "what-if" cambiando variables road_defect o traffic_control_device en el modelo predictivo*	*Priorizar inversión en infraestructura y señalización según impacto estimado en reducción de accidentes graves*
5. ¿Cuál es el número esperado de accidentes graves por día/mes y cómo priorizar recursos preventivos?	*Predicción agregada de accidentes graves por día/mes, ranking de tramos viales por riesgo*	*Planificación operativa de patrullaje, mantenimiento y campañas educativas según pronóstico de riesgo*

2. Contexto y problemática

*La **Municipalidad de Ciudad X** ha experimentado un aumento sostenido en accidentes de tráfico en los últimos años, especialmente accidentes graves que involucran lesiones incapacitantes o fatales. Este fenómeno genera **costes humanos y económicos elevados**, incluyendo:*

- *Atención médica de urgencia y hospitalización.*
- *Daños a la infraestructura vial y vehículos.*
- *Interrupciones de tráfico que afectan la productividad y el transporte público.*
- *Costes legales y de seguros.*

Actualmente, las decisiones sobre inspecciones viales, patrullaje policial y campañas de seguridad se toman de manera reactiva, es decir, después de que ocurren los accidentes, sin poder anticipar zonas o franjas horarias de alto riesgo.

***Problema clave:** La ciudad carece de un sistema predictivo que permita **identificar proactivamente los accidentes graves** y priorizar acciones preventivas.*

Problema:

- *Los accidentes de tráfico representan un riesgo significativo para la seguridad vial y generan costos humanos y económicos considerables.*
- *Las autoridades y empresas de transporte necesitan priorizar la atención a los accidentes más graves.*
- *El desafío consiste en predecir la gravedad de un accidente (leve o grave) a partir de datos históricos de incidentes, características de la vía, condiciones climáticas y factores del accidente.*
- *El dataset presenta desbalance de clases, ya que los accidentes graves representan aproximadamente el 20% del total.*

Solución:

- *Desarrollo de un modelo de machine learning supervisado capaz de clasificar la gravedad de los accidentes.*
- *Uso de técnicas de preprocesamiento, como generación de variables temporales y codificación de variables categóricas.*
- *Comparación de múltiples algoritmos de clasificación:*
  - *Decision Tree, Random Forest, Gradient Boosting, XGBoost.*
  - *Regresión Logística con balanceo de clases y SMOTE.*
  - *Algoritmos específicos para clases desbalanceadas: Balanced Random Forest y Easy Ensemble.*
- *Selección del mejor modelo basado en métricas clave (F1-score, ROC-AUC, Average Precision).*
- *Despliegue del modelo seleccionado mediante FastAPI para permitir predicciones en tiempo real a partir de inputs de accidentes.*

Enfoque:

- *Exploración y análisis del dataset para entender distribución, valores faltantes y relaciones entre variables.*
- *No se gestionaron outliers, ya que la eliminación afectaría la clase minoritaria (accidentes graves).*
- *Implementación de validación cruzada para evaluar estabilidad y generalización de los modelos:*
  - *GridSearchCV para Logistic Regression y Random Forest.*
  - *StratifiedKFold para Balanced Random Forest y Easy Ensemble.*
- *Métricas principales consideradas:*
  - *Precision, Recall, F1-score, ROC-AUC y Average Precision, enfocándose en la detección de accidentes graves.*
- *Selección del Balanced Random Forest como modelo final debido a su balance entre precisión y recall para la clase minoritaria.*

Conclusiones

- *Los modelos de ensamble balanceados (Balanced RF, Easy Ensemble) muestran mejor capacidad para detectar accidentes graves en datasets desbalanceados.*
- *La regresión logística con balance de clases o SMOTE mejora el recall de la clase minoritaria pero puede reducir precisión general.*
- *Random Forest y Gradient Boosting funcionan bien con la clase mayoritaria, pero requieren técnicas adicionales para manejar el desbalance.*
- *La validación cruzada confirmó la estabilidad y generalización del Balanced RF.*
- *Lessons Learned:*
- *El manejo del desbalance es crítico para modelos de clasificación de accidentes.*
- *La selección de métricas adecuadas (AP, F1-score) es esencial para interpretar correctamente el desempeño.*
- *La implementación de un API permite llevar el modelo a un entorno práctico de predicción en tiempo real, facilitando decisiones preventivas y de gestión de riesgos.*

3. Datos proporcionados por el cliente/ciudad

*Los datos históricos de accidentes contienen **información detallada sobre las condiciones del accidente**, incluyendo:*

Variable	***Descripción***	***Tipo de dato***
crash_date	*Fecha del accidente*	*datetime*
crash_hour	*Hora del accidente*	*int*
crash_day_of_week	*Día de la semana*	*categórica*
crash_month	*Mes*	*categórica*
traffic_control_device	*Dispositivo de control de tráfico involucrado*	*categórica*
weather_condition	*Condiciones climáticas*	*categórica*
lighting_condition	*Condiciones de iluminación*	*categórica*
first_crash_type	*Tipo inicial del accidente*	*categórica*
trafficway_type	*Tipo de vía*	*categórica*
alignment	*Alineación de la vía*	*categórica*
roadway_surface_cond	*Estado de la superficie*	*categórica*
road_defect	*Defectos en la vía*	*categórica*
crash_type	*Tipo global del accidente*	*categórica*
intersection_related_i	*Si ocurrió en intersección*	*binaria*
damage	*Extensión del daño*	*categórica*
prim_contributory_cause	*Causa principal del accidente*	*categórica*
num_units	*Número de vehículos involucrados*	*numérica*
most_severe_injury	*Lesión más grave*	*categórica*
injuries_total	*Total de lesiones*	*numérica*
injuries_fatal	*Número de lesiones fatales*	*numérica*
injuries_incapacitating	*Lesiones incapacitantes*	*numérica*
injuries_non_incapacitating	*Lesiones no incapacitantes*	*numérica*
injuries_reported_not_evident	*Lesiones reportadas no evidentes*	*numérica*
injuries_no_indication	*Sin indicio de lesión*	*numérica*

*Esta información **permite modelar la probabilidad de que un accidente sea grave** según múltiples factores combinados (hora, tipo de vía, clima, control de tráfico, etc.).*

4. Beneficio esperado:

*Con estos datos se pueden construir modelos de **predicción de accidentes graves**, que permitirán:*

- *Priorizar inspecciones y mantenimientos en zonas de alto riesgo.*
- *Ajustar patrullaje policial y señalización preventiva.*
- *Diseñar campañas educativas dirigidas a los horarios y vías de mayor riesgo.*
- *Reducir la frecuencia y gravedad de accidentes, optimizando recursos públicos.*

Beneficios específicos para la ciudad:

- ***Reducción de accidentes graves:** a través de intervenciones preventivas focalizadas.*
- ***Optimización de recursos públicos:** mejor distribución de patrullaje, inspección vial y mantenimiento.*
- ***Decisiones basadas en datos:** priorización objetiva y reproducible.*
- ***Mejora de la seguridad ciudadana** y fortalecimiento de la reputación de la administración.*
- ***Potencial para futuras ampliaciones:** integración con sistemas de tráfico en tiempo real o IoT vial para generar alertas dinámicas.*

5. Alcance del proyecto

***Datos utilizados:** variables descritas en el dataset (fecha, hora, tipo de vía, clima, iluminación, defectos, causa, número de vehículos, severidad de lesiones, etc.).*

Salida esperada:

- *Modelo predictivo que, dado un tramo vial y franja horaria, **calcule la probabilidad de accidente grave**.*
- *Dashboard de visualización con **mapas de riesgo y métricas de severidad**.*
- *Informe con **recomendaciones de priorización de recursos** para prevención de accidentes.*

6. Posibles acciones basadas en los resultados

- ***Patrullaje vial focalizado:** enviar recursos a zonas/hora de mayor riesgo.*
- ***Inspecciones y mantenimiento vial:** priorizar reparaciones donde la superficie o defectos elevan riesgo.*
- ***Campañas educativas:** horarios y vías críticas según patrones detectados.*
- ***Alertas tempranas:** integración futura con sistemas de tráfico en tiempo real.*