**1. Estudios sobre predicción de duración/tiempo de incidentes**

| **Autor / Año** | **Dataset** | **Variables usadas** | **Modelos probados** | **Métricas reportadas** | **Resultados clave** |
| --- | --- | --- | --- | --- | --- |
| Grigorev (2024) | Accidentes de tráfico en Sídney (datasets públicos + reportes) | Hora, ubicación, tipo de accidente, condiciones del tráfico, clima | Random Forest, XGBoost, LightGBM, CatBoost | MAE, RMSE | Boosting (XGBoost/LightGBM) superó a RF y regresión. MAE ~15–20 min. Problemas de outliers. |
| Chen et al. (2024) | Accidentes en autopistas chinas | Datos multimodales: estructurados (hora, clima, tráfico), sensores, texto de reportes | Ensemble deep learning (CNN + GRU + atención) | MAE, RMSE, MAPE | Modelo multimodal redujo error 12–15% frente a métodos clásicos. Texto (NLP) añadió valor. |
| Corbally et al. (2024) | Incidentes en autopistas (Reino Unido) | Ubicación, hora, tipo de incidente, recursos desplegados | SVM, Random Forest, Gradient Boosting | MAE, RMSE | GBM logró mejor desempeño; RF fue más robusto con outliers. |

**Estudios en contexto de emergencias médicas (EMS / ED)**

| **Autor / Año** | **Dataset** | **Variables usadas** | **Modelos probados** | **Métricas reportadas** | **Resultados clave** |
| --- | --- | --- | --- | --- | --- |
| Hill et al. (2025) | Datos de EMS (respuesta de ambulancias, Europa) | Hora, ubicación, tráfico, disponibilidad de ambulancias | Random Forest, Gradient Boosting, regresión lineal | MAE, RMSE | RF y GBM lograron menor error; importancia de disponibilidad de recursos como variable crítica. |
| Ricciardi et al. (2024) | Emergencias hospitalarias (Italia) | Demografía paciente, motivo consulta, hora, disponibilidad ED | RF, SVM, KNN, regresión | MAE, RMSE | RF superó en desempeño. MAE < 20 min en tiempos de espera ED. |
| Wang et al. (2025) | Datos de hospitales en EE. UU. | Variables administrativas + clínicas + tiempo de registro | XGBoost + SHAP | Fairness, MAE, RMSE | Buen desempeño (MAE < 30 min), pero detectaron sesgos entre subgrupos (género, etnia). |

**Estudios en aseguradoras / industria**

| **Autor / Año** | **Dataset** | **Variables usadas** | **Modelos probados** | **Métricas reportadas** | **Resultados clave** |
| --- | --- | --- | --- | --- | --- |
| Shift Technology (2025) | Datos de siniestros en aseguradoras (caso de negocio, no público) | Ubicación, hora, tipo de accidente, disponibilidad de gestores | AutoML (con GBM, RF, redes) | MAE, SLA internos | Enfocado en uso en tiempo real para priorización de reclamos; destaca la importancia de integración con sistemas de gestión. |
| Budiman et al. (2023) | Datos de urgencias hospitalarias (similar a aseguradoras en operación) | Variables operativas y de ocupación | Random Forest ensembles | RMSE, MAE | Ensembles de RF mostraron robustez frente a datos desbalanceados. |

**Variables predictoras recurrentes**  
Los estudios coinciden en que las variables con mayor aporte explicativo incluyen: 1. 1.

1. Ubicación geográfica (coordenadas o zona) (Destino)
2. Ubicación geográfica (coordenadas o zona) (Origen) podemos reconstruir esta información
3. timestamp (hora/día/festivo)
4. Condiciones meteorológicas
5. Nivel de tráfico/velocidad media en la vía
6. Tipo y severidad del accidente
7. Recursos de apoyo, y la disponibilidad/ocupación de unidades en campo.

**Problemas de datos y retos metodológicos**  
Los obstáculos más citados son: calidad de datos (registros incompletos o inconsistentes), valores atípicos extremos (incidentes con duraciones muy largas), truncamiento de tiempos, y distribuciones fuertemente sesgadas (muchos eventos breves y pocos muy largos). La heterogeneidad de fuentes (registros administrativos, texto libre) implica además problemas de *feature engineering(*hacer que los datos sean más comprensibles*)*. Por ello, gran parte de la literatura dedica secciones largas a limpieza, imputación, tratamiento de outliers y validación temporal (evitar fugas de información)

# **Observaciones transversales**

* **Modelos ganadores:** en la mayoría de casos **Random Forest y Gradient Boosting** ofrecen mejor trade-off entre precisión e interpretabilidad; en datos multimodales, **deep learning híbrido (CNN/GRU)** supera a los métodos clásicos.
* **Variables más influyentes:** disponibilidad de recursos (ambulancias/gestores), ubicación y tiempo del día, condiciones externas (clima, tráfico).
* **Problemas comunes:** outliers, registros incompletos y distribuciones sesgadas; se resolvieron con imputación, winsorización o validación temporal.
* **Novedad:** incorporación de texto (NLP) y fairness/interpretabilidad (SHAP) en los modelos más recientes.