

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Práctica 2: Limpieza y Análisis de Datos

Máster en Ciencia de Datos

Universitat Oberta de Catalunya

Análisis de la Base de Datos Nacional de Colisiones de Canadá (2015-2020)

Creado by : ABDALLAH TEGGUER

Enlace al repositorio Git:
https://github.com/Abdallahtegguer/prac2_tcvd

Fecha de entrega: 7 de enero de 2026

1. Descripción del Dataset

Este trabajo analiza la Base de Datos Nacional de Colisiones (National Collision Database - NCDB) de Transport Canada. Este dataset contiene información detallada sobre todas las colisiones de tráfico reportadas por la policía en Canadá, siendo una de las fuentes más completas sobre seguridad vial del país.

Importancia del dataset: La seguridad vial es un tema de gran relevancia social y económica. Comprender los factores que influyen en la severidad de los accidentes puede ayudar a desarrollar políticas de prevención más efectivas y salvar vidas.

Pregunta de investigación: ¿Qué factores determinan si un accidente de tráfico resulta en fatalidades versus solo heridos? Buscamos identificar patrones temporales, demográficos y situacionales asociados a la severidad de las colisiones.

Variables principales del dataset:

Variable	Descripción	Tipo
C_YEAR	Año de la colisión	Numérica
C_MNTH	Mes de la colisión (1-12)	Numérica
C_WDAY	Día de la semana (1=Lun a 7=Dom)	Numérica
C_HOUR	Hora de la colisión (0-23)	Numérica
C_SEV	Severidad (1=Fatal, 2=Heridos)	Categórica
C_VEHIS	Número de vehículos	Numérica
C_WTHR	Condiciones climáticas	Categórica
P_SEX	Sexo de la persona (M/F)	Categórica
P_AGE	Edad de la persona	Numérica
V_YEAR	Año del vehículo	Numérica

Tamaño del dataset: Se integraron datos de 6 años (2015-2020), con un total de 1,664,553 registros iniciales y 23 variables. Tras la limpieza: 1,646,796 registros y 28 variables.

2. Integración y Selección de Datos

Se integraron seis archivos CSV correspondientes a los años 2015-2020. La elección de este rango temporal permite un análisis robusto con datos recientes sin comprometer los recursos computacionales.

Año	Registros	Porcentaje
2015	312,042	18.75%
2016	306,260	18.40%
2017	289,823	17.41%
2018	285,382	17.15%
2019	272,301	16.36%
2020	198,745	11.94%
Total	1,664,553	100%

Nota: La reducción en 2020 es consistente con las restricciones de movilidad por la pandemia COVID-19.

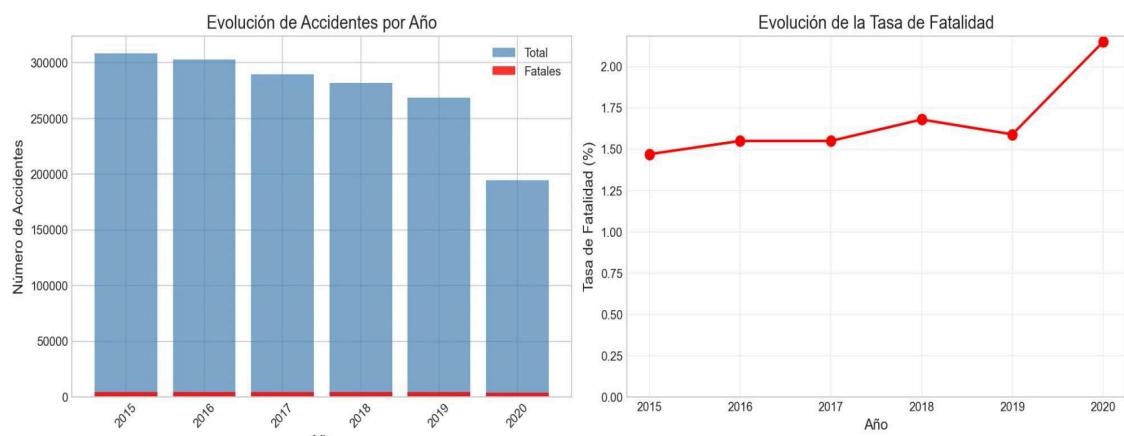


Figura 1: Evolución temporal de accidentes y tasa de fatalidad (2015-2020)

3. Limpieza de Datos

3.1 Gestión de Valores Faltantes

El dataset NCDB utiliza códigos especiales para valores desconocidos ('U', 'UU', 'QQ', 'N', 'NN'). Estos fueron convertidos a NaN para su tratamiento sistemático.

Variable	Nulos	%
P_SAFE	374,182	22.48%
C_CONF	143,954	8.65%
P_AGE	113,976	6.85%
P_SEX	100,920	6.06%
V_TYPE	93,869	5.64%

Estrategia de imputación:

- Variables numéricas (P_AGE, V_YEAR, C_HOUR): Imputación con mediana
- Variables categóricas con <10% nulos: Imputación con moda
- Variables categóricas con >10% nulos: Categoría "Desconocido"
- Registros con múltiples nulos críticos: Eliminación (17,685 registros, 1.06%)

3.2 Tipos de Datos

Se realizaron las conversiones apropiadas: variables numéricas a int64/float64, categóricas al tipo 'category', y se creó la variable objetivo binaria FATAL (1=Fatal, 0=No fatal). Distribución: 98.37% No Fatal, 1.63% Fatal.

3.3 Valores Extremos

Se aplicó el método IQR para identificar outliers. Tratamiento aplicado:

- P_AGE: Limitado al rango [0, 110] años
- C_VEHIS: Capping a máximo 20 vehículos (1,127 registros afectados)
- V_YEAR: Años inválidos reemplazados con mediana (2010)

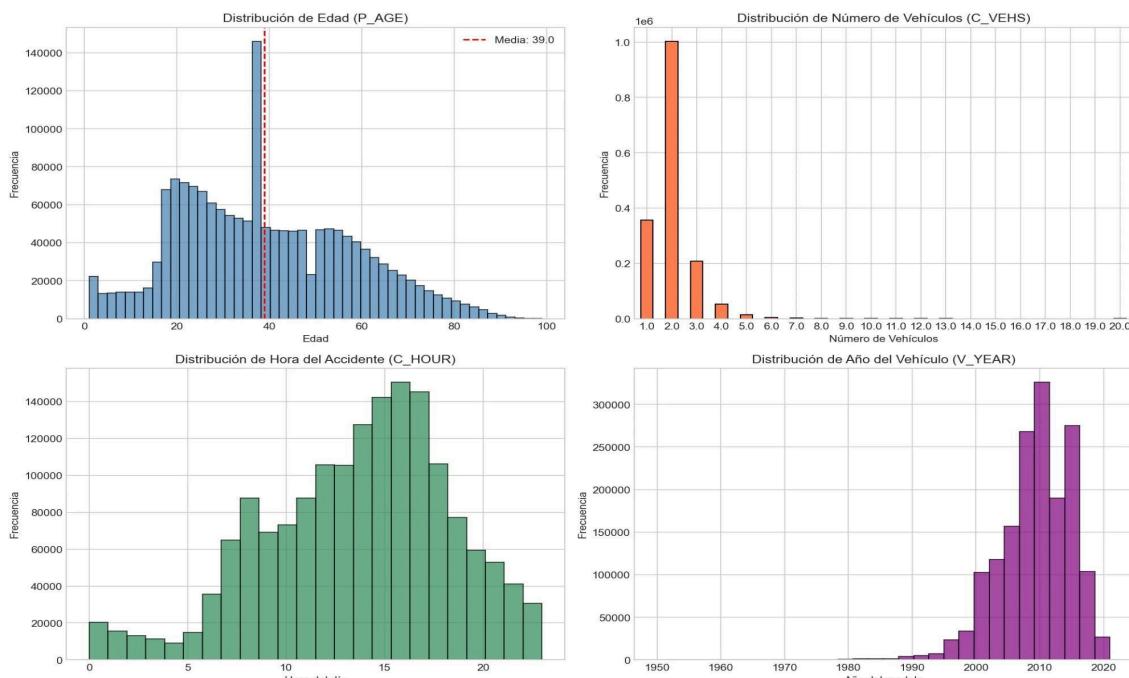


Figura 2: Distribución de variables numéricas después de la limpieza

3.4 Otros Métodos de Limpieza

- Eliminación de 72 duplicados exactos
- Estandarización de valores categóricos (P_SEX a mayúsculas)
- Creación de variables derivadas: VEHICLE_AGE (antigüedad, media=8 años), TIME_PERIOD (Madrugada/Mañana/Tarde/Noche), WEEKEND (indicador fin de semana), AGE_GROUP (grupos de edad)

Resumen: Registros finales: 1,646,796 (99% de los originales). Variables finales: 28.

4. Análisis de los Datos

4.1 Modelo Supervisado: Random Forest Classifier

Se implementó un modelo Random Forest para predecir la fatalidad de accidentes.

Configuración: n_estimators=100, max_depth=10, class_weight='balanced'

División: 70% entrenamiento (1,152,631), 30% test (493,986)

Métrica	Valor
Accuracy	74.04%
Recall (Fatal)	52%
Precision (Fatal)	3%
F1-Score (Fatal)	0.06

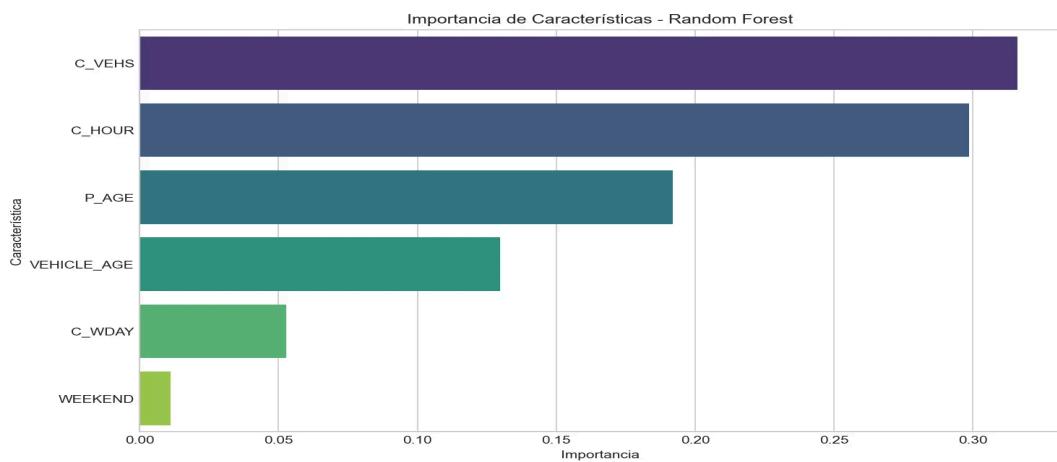


Figura 3: Importancia de características del modelo Random Forest

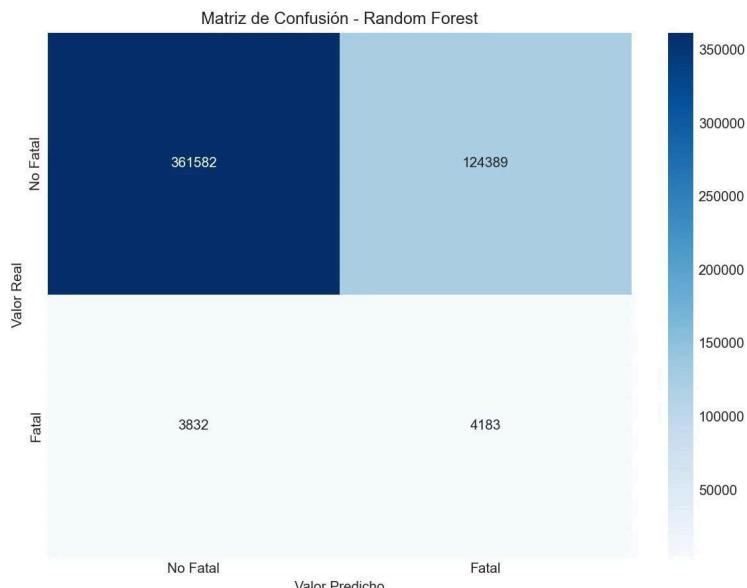


Figura 4: Matriz de confusión del modelo

4.1.2 Modelo No Supervisado: K-Means Clustering

Se aplicó K-Means para identificar grupos naturales en los datos. Usando el método del codo y Silhouette Score, se determinó K=4 como número óptimo de clusters.

Cluster	Tamaño	Edad	Hora	Tasa Fatal
0 - Jóvenes tarde	30.9%	27.4	17:00	1.20%
1 - Mayores día	23.2%	61.2	14:00	1.87%
2 - Adultos mañana	20.3%	33.6	08:00	1.65%
3 - Fin de semana	25.6%	37.8	14:00	1.90%

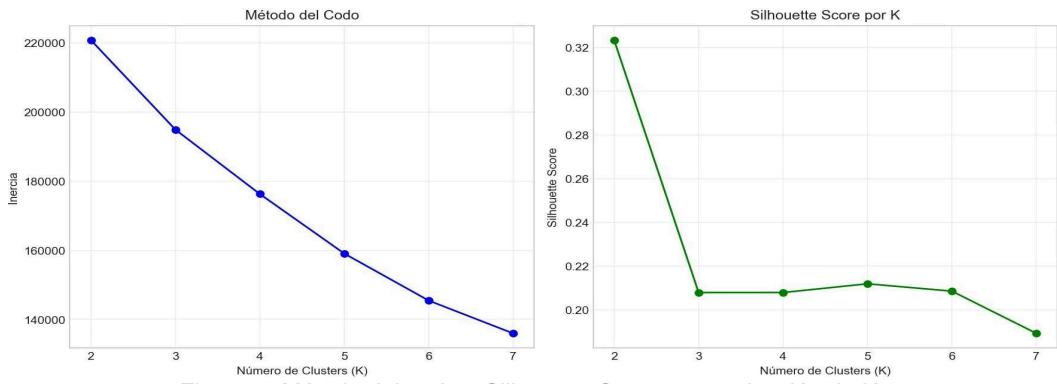


Figura 5: Método del codo y Silhouette Score para selección de K

4.2 Contraste de Hipótesis

Verificación de supuestos:

- Normalidad (Shapiro-Wilk): Se rechaza para todas las variables ($p < 0.05$)
- Homocedasticidad (Levene): Varianzas no homogéneas ($p < 0.05$)
- Conclusión: Se utilizan pruebas no paramétricas

Hipótesis 1: Diferencia de edad entre accidentes fatales y no fatales

- H_0 : No hay diferencia significativa en la edad promedio
- Test: Mann-Whitney U (no paramétrico)
- Resultados: Edad fatal = 41.77 años vs No fatal = 38.92 años
- $U = 23,360,237,741.50$, p-value = 1.80e-92
- **Conclusión: Se RECHAZA H_0 . Existe diferencia significativa.**

Hipótesis 2: Relación entre período del día y severidad

- H_0 : No hay relación entre el período del día y la severidad
- Test: Chi-cuadrado de independencia
- Resultados: $\chi^2 = 3,666.59$, gl = 3, p-value < 0.001
- Tasas: Madrugada (3.42%) > Noche (1.74%) > Mañana (1.45%) > Tarde (1.33%)
- **Conclusión: Se RECHAZA H_0 . El período del día influye en la severidad.**

Hipótesis 3: Relación entre fin de semana y accidentes fatales

- H_0 : La proporción de fatales es igual entre semana y fin de semana
- Test: Chi-cuadrado
- Resultados: $\chi^2 = 317.58$, gl = 1, p-value = 4.88e-71
- Tasas: Fin de semana (1.94%) vs Entre semana (1.53%)
- **Conclusión: Se RECHAZA H_0 . Los fines de semana tienen mayor tasa de fatalidad.**

5. Representación de Resultados

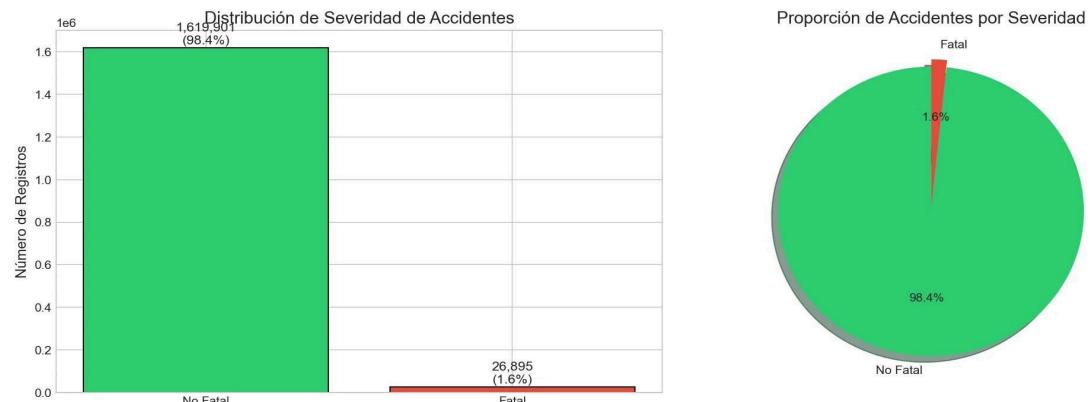


Figura 6: Distribución de severidad de accidentes (Fatal vs No Fatal)

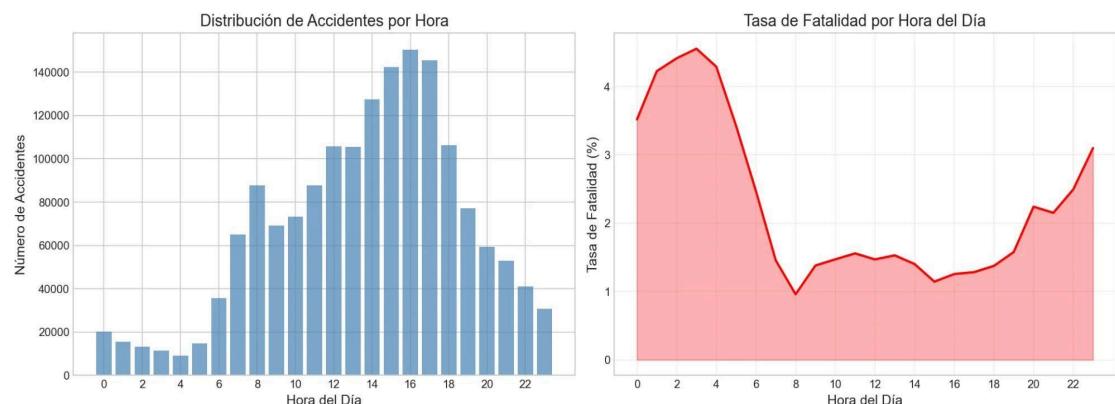


Figura 7: Distribución de accidentes y tasa de fatalidad por hora del día

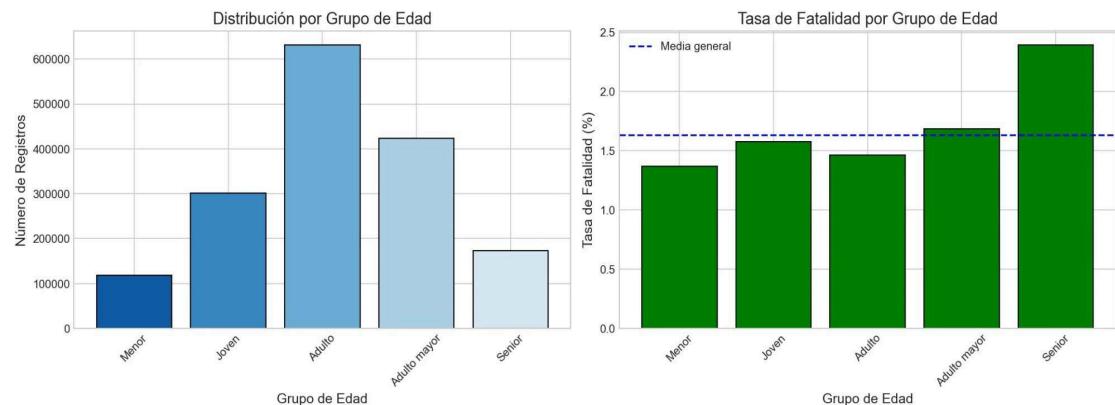


Figura 8: Distribución y tasa de fatalidad por grupo de edad

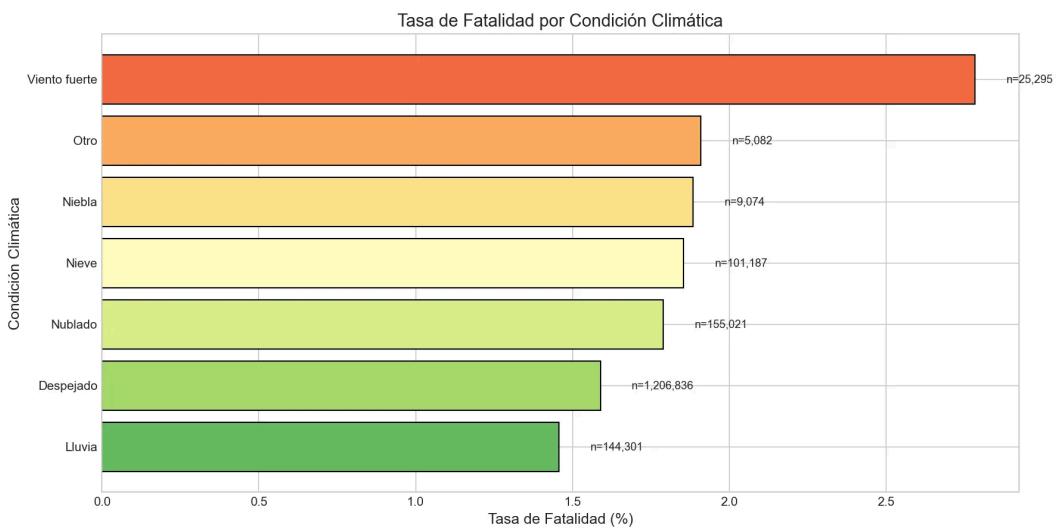


Figura 9: Tasa de fatalidad por condición climática

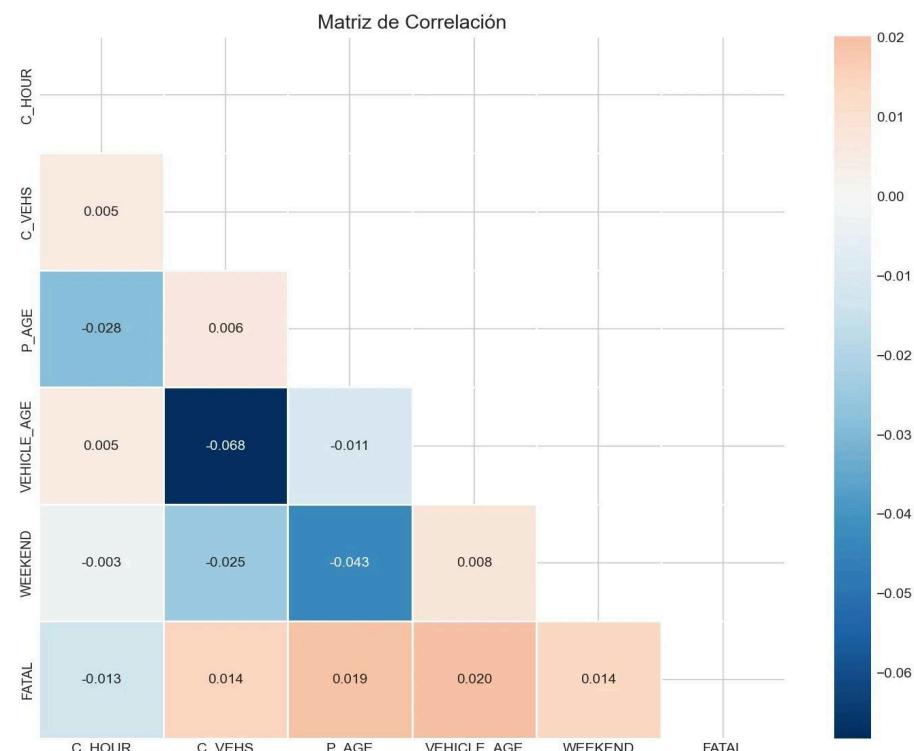


Figura 10: Matriz de correlación entre variables numéricas

6. Resolución del Problema - Conclusiones

Hallazgos principales:

1. Factores temporales:

- La madrugada (00:00-06:00) presenta la mayor tasa de fatalidad (3.42%), más del doble que otros períodos
- Los fines de semana tienen una tasa significativamente mayor (1.94% vs 1.53%)

2. Factores demográficos:

- Las personas mayores tienen mayor riesgo de fatalidad
- Edad promedio en accidentes fatales: 41.77 años vs 38.92 en no fatales
- El grupo Senior (65+) presenta la tasa más alta (2.4%)

3. Factores del vehículo y condiciones:

- El número de vehículos es el predictor más importante (importancia: 0.316)
- El viento fuerte presenta la mayor tasa de fatalidad por condición climática
- La antigüedad del vehículo también contribuye al riesgo

4. Respuesta a la pregunta de investigación:

Los principales factores que determinan la fatalidad en accidentes de tráfico son:

1. Hora del accidente (madrugada = mayor riesgo)
2. Número de vehículos involucrados
3. Edad de las personas (mayor edad = mayor riesgo)
4. Día de la semana (fin de semana = mayor riesgo)
5. Condiciones climáticas adversas

5. Limitaciones:

- Fuerte desbalanceo de clases (solo 1.63% de accidentes fatales)
- El modelo presenta bajo precision para la clase Fatal debido al desbalanceo
- No se dispone de información sobre consumo de alcohol o uso de cinturón en todos los casos
- Datos de 2020 pueden estar influenciados por la pandemia COVID-19

7. Código

El código completo está disponible en el repositorio Git. El archivo principal es **practica2_analisis.py** y contiene:

- Carga e integración de datos (6 archivos CSV)
- Limpieza y preprocesamiento completo
- Análisis exploratorio de datos
- Implementación de Random Forest Classifier
- Implementación de K-Means Clustering
- Contraste de hipótesis (Mann-Whitney U, Chi-cuadrado)
- Generación de todas las visualizaciones

Librerías utilizadas: pandas, numpy, matplotlib, seaborn, scipy, scikit-learn

Estructura del repositorio:

- README.md
- practica2_analisis.py
- *.csv & *.xlsx originales (1999-2020)
- graficos(.png)
- memoria_practica2.pdf

8. Vídeo

El vídeo explicativo de la práctica está disponible en:

Tabla de Contribuciones

La siguiente tabla indica la participación de cada integrante en los diferentes aspectos del trabajo:

Contribuciones	Firma
Investigación previa	abdallah tegguer
Redacción de las respuestas	abdallah tegguer
Desarrollo del código	abdallah tegguer
Participación en el vídeo	abdallah tegguer

Uso de inteligencia artificial

En el desarrollo de este proyecto se ha utilizado de forma puntual y responsable una herramienta de inteligencia artificial como apoyo auxiliar.

Su uso se limitó a tareas de estructuración del proyecto, revisión de coherencia, detección de posibles inconsistencias y mejora de la redacción académica en español.

La herramienta también se empleó al final del proceso para verificar la completitud global de la memoria desde una perspectiva organizativa.

En ningún caso la inteligencia artificial intervino en el desarrollo del código, el análisis de los datos ni en la obtención de resultados, los cuales han sido realizados íntegramente por el autor.

El uso de esta herramienta tuvo como único objetivo optimizar el proceso de trabajo, mejorar la organización y garantizar una presentación más clara y profesional del proyecto.

Justificación de la selección del conjunto de datos

Para el desarrollo de la Práctica II se ha utilizado un subconjunto del conjunto de datos original presentado en la Práctica I, concretamente el correspondiente al período comprendido entre **2015 y 2020**. Esta decisión responde a **limitaciones técnicas derivadas del volumen total de datos**, que impedían una ejecución eficiente y estable del código en el entorno de trabajo disponible.

El uso del conjunto de datos completo implicaba **tiempos de ejecución excesivamente elevados** y problemas de rendimiento durante las fases de implementación, prueba y validación del código, lo que dificultaba el desarrollo iterativo y la correcta evaluación de los resultados.

Por este motivo, se optó por trabajar con los datos más recientes, garantizando un volumen manejable y representativo que permitiera **implementar, ejecutar y validar correctamente el flujo de trabajo**, sin comprometer los objetivos metodológicos ni la validez del análisis realizado en esta práctica.