CENTRO POLITÉCNICO SUPERIOR
MALVINAS ARGENTINAS

Tercer Entrega

Informe Final

Para este trabajo se utilizó un dataset de accidentes de tránsito del **Municipio de Río Grande** correspondiente al período **2009-2023**.

Los datos fueron obtenidos de registros oficiales y contienen información mensual sobre la cantidad de accidentes, víctimas, lesionados, distribución por zona de ocurrencia (urbana y rural) y casos con alcohol positivo. (Archivo en formato excel).

El objetivo es analizar estos datos y construir modelos predictivos para anticipar la cantidad de accidentes con alcohol positivo y el número de accidentes por zona.

Para comenzar con la exploración del dataset se convirtió a formato csv, se realizó una limpieza y estandarización de los datos donde las primeras observaciones fueron que el número de accidentes totales de casos de alcohol positivo muestra variaciones a los largo de los años, dándose los mayores caso positivos en algunos meses en particular.

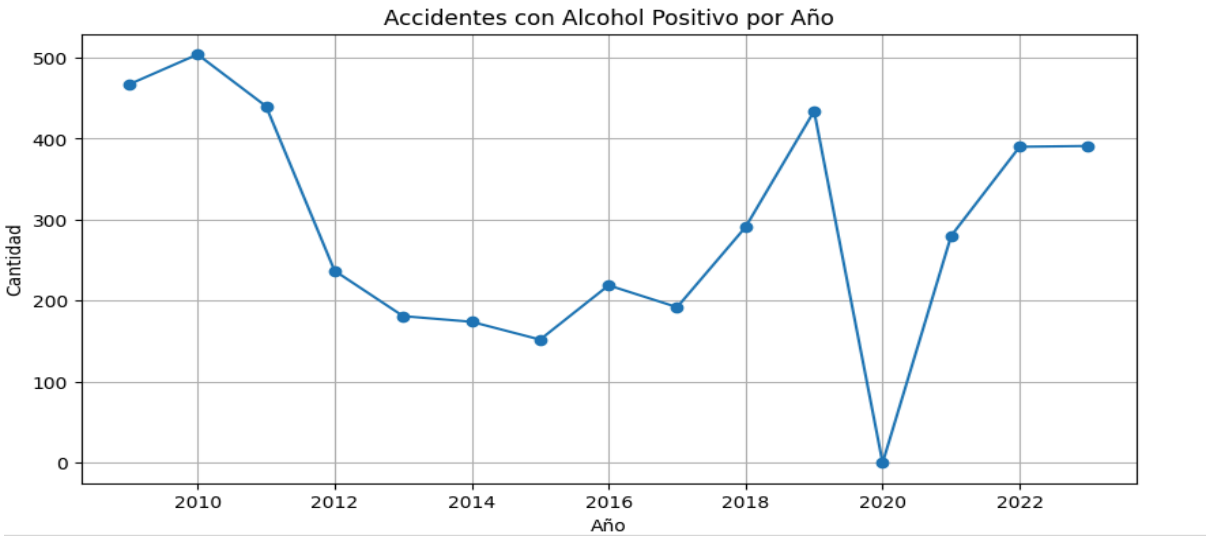
También se pudo observar que la gran parte de accidentes ocurren en zona urbana y los casos de alcohol positivo no dependen de la cantidad de accidentes.

se incluyeron gráficos :

- ➢ Accidentes con alcohol positivo por año y mes.
- ➢ Accidentes por zona (urbana/rural) por año.
- ➢ Matriz de correlación.
- ➢ Resultados de predicción (real vs predicho).

Accidentes con alcohol por año

```
[ ]  plt.figure(figsize=(10, 5))
     df.groupby('Año')['Alcohol_positivo'].sum().plot(marker='o')
     plt.title('Accidentes con Alcohol Positivo por Año')
     plt.xlabel('Año')
     plt.ylabel('Cantidad')
     plt.grid()
     plt.show()
```

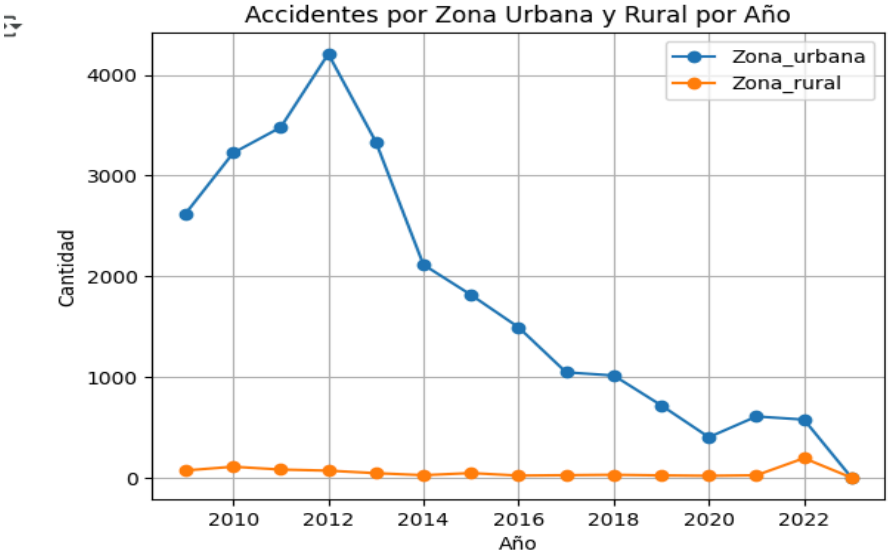


Se utilizó un gráfico de líneas para analizar los accidentes con alcohol positivo por año donde en el año 2009 al 2011 el número de accidentes con

alcohol positivo fue muy alto, alcanzando un máximo en 2010 (más de 500 casos). A partir de 2011 se observa una disminución en la cantidad de accidentes de este tipo, tocando mínimos en 2014 y 2015.Desde 2016 hasta 2019 la cantidad de casos aumenta gradualmente donde va variando año a año, en el año 2020 se observa un descenso a casi cero este dato se debe a la pandemia de COVID-19 y las restricciones de circulación, lo cual redujo la movilidad y la exposición a situaciones de riesgo vial., ya en en el año 2021 los casos vuelven a aumentar y se estabilizan a 390 y 400 casos anuales para 2022 y 2023.

Accidentes por zona urbana y rural

```
[26] plt.figure(figsize=(10, 5))
     df.groupby('Año')[['Zona_urbana', 'Zona_rural']].sum().plot(marker='o')
     plt.title('Accidentes por Zona Urbana y Rural por Año')
     plt.xlabel('Año')
     plt.ylabel('Cantidad')
     plt.grid()
     plt.show()
```

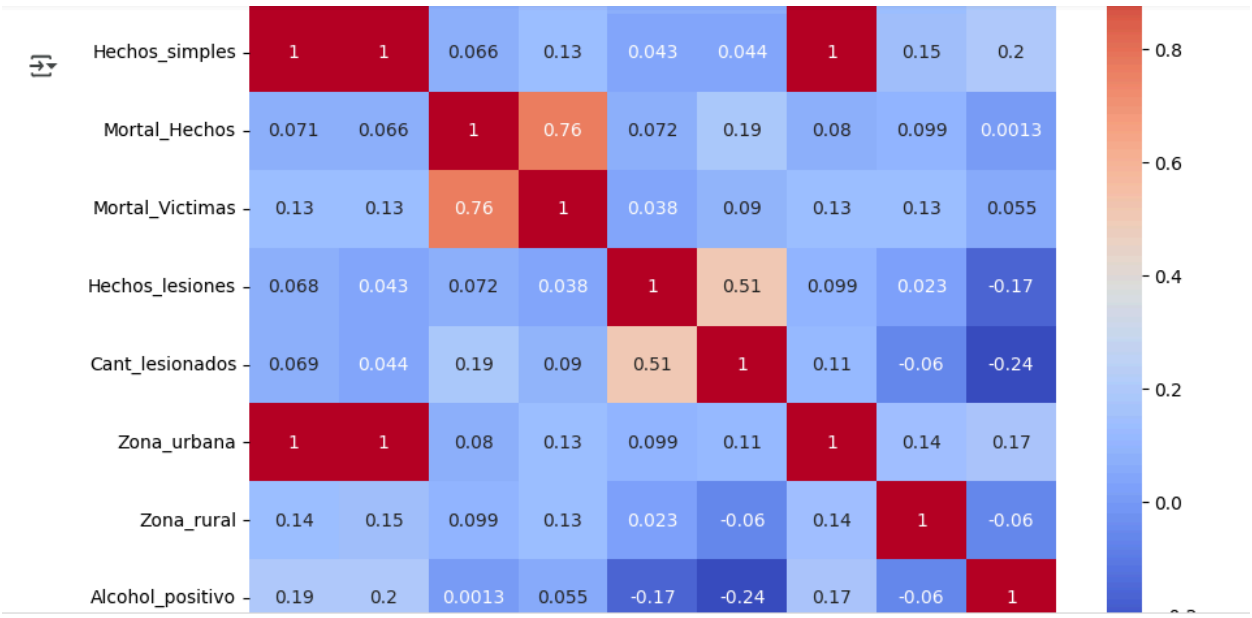


El gráfico muestra la evolución anual de accidentes de tránsito en zona urbana y rural , donde la zona urbana presenta la mayor cantidad de accidentes en todo el periodo analizado a diferencia de la zona rural que son muy pocos y estables.

En relación a la zona urbana los mayores accidentes se dieron en el año 2012 superando los 4000,con una baja notable en los años siguientes.

Correlacion entre variables numericas

```
import seaborn as sns
plt.figure(figsize=(10, 7))
sns.heatmap(df[cols_to_numeric].corr(), annot=True, cmap='coolwarm')
plt.title('Matriz de Correlación')
plt.show()
```



En la matriz muestra una correlación entre “Hechos simples” y “zona Urbana” (1.0) donde se demuestra que la gran parte de los accidentes se dan en dicha zona.

Fuerte correlación entre “Mortal Hechos” y “Mortal Víctimas” (0.76), lo cual es esperable ya que ambos se refieren a la gravedad de los accidentes.

Correlación moderada entre “Hechos lesiones” y “Cant lesionados” (0.51), mostrando que en general, más hechos con lesiones implican más lesionados, aunque no siempre en proporción exacta.

Baja correlación entre “Alcohol positivo” y el resto de las variables (en general menores a 0.2), indicando que los accidentes con alcohol positivo son un fenómeno que no depende directamente del volumen total de accidentes, ni de la gravedad o la zona. Esto hace que su predicción sea un desafío, reforzando la necesidad de datos adicionales para mejorar el modelo.

Objetivos de Modelado: Se plantearon dos objetivos de predicción.

- ➢ Predecir la cantidad de accidentes con alcohol positivo por mes y año.
- ➢ Predecir el número de accidentes por zona urbana.

Predicción de accidentes con alcohol positivo (Modelo1)

Algoritmos utilizados: Regresión Lineal y Random Forest Regressor (modelo principal).

Variables empleadas: Año, mes, total de accidentes, accidentes en zona urbana y rural, cantidad de lesionados.

Ajuste de hiper parámetros: Se utilizaron los valores por defecto de Random Forest para una primera aproximación.

Preparacion de los datos

```
[ ]  from sklearn.model_selection import train_test_split
     from sklearn.ensemble import RandomForestRegressor
     from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

Modelo-1 Prediccion de Alcohol Positivo

```
[ ]  features_1 = ['Año', 'Mes_num', 'Total', 'Zona_urbana', 'Zona_rural', 'Cant_lesionados']
     X1 = df[features_1]
     y1 = df['Alcohol_positivo']

     X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1, test_size=0.2, random_state=42)
```

***Modelo de Regresion Lineal ***

```
[ ]  lr1 = LinearRegression()
     lr1.fit(X1_train, y1_train)
     y1_pred_lr = lr1.predict(X1_test)
```

Modelo Ramdon Fores

```
[ ]  rf1 = RandomForestRegressor(random_state=42)
     rf1.fit(X1_train, y1_train)
     y1_pred_rf = rf1.predict(X1_test)
```

***Evaluacion ***

***Evaluacion ***

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

def print_metrics(y_test, y_pred, modelo):
    print(f"\n--- Métricas {modelo} ---")
    print("MAE:", mean_absolute_error(y_test, y_pred))
    # Calculate RMSE manually by taking the square root of MSE
    print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
    print("R²:", r2_score(y_test, y_pred))

print_metrics(y1_test, y1_pred_lr, "Regresión Lineal (Alcohol)")
print_metrics(y1_test, y1_pred_rf, "Random Forest (Alcohol)")
```

```
--- Métricas Regresión Lineal (Alcohol) ---
MAE: 11.59295402747763
RMSE: 14.336960551740583
R²: -0.0021321455901899267

--- Métricas Random Forest (Alcohol) ---
MAE: 9.543611111111112
RMSE: 12.604918462427452
R²: 0.22537609696641403
```

importancia de ramdo Fores

```
importances = rf1.feature_importances_
for col, imp in zip(features_1, importances):
    print(f"Importancia {col}: {imp:.3f}")
```

```
Importancia Año: 0.500
Importancia Mes_num: 0.066
Importancia Total: 0.108
Importancia Zona_urbana: 0.110
Importancia Zona_rural: 0.127
Importancia Cant_lesionados: 0.088
```

Random Forest superó a la regresión lineal en ambos casos,especialmente en la predicción de accidentes por zona urbana.
En la predicción de alcohol positivo, el modelo tuvo precisión aceptable, aunque limitada por la baja correlación con otras variables.

Un R 2 negativo indica que el modelo de regresión lineal no logra explicar la variabilidad de los datos, osea que la regresión lineal no es un modelo adecuado para este problema con las variables actuales.

Se utilizó " La importancia de variables" del modelo Random Forest para saber cuáles son las variables que más influyen en la predicción del modelo planteado por ejemplo en la variable "Año" es el factor más relevante en la predicción de accidentes con alcohol positivo, seguida por la "Zona rural" y la "Zona urbana". Esto sugiere que existen factores temporales y geográficos que impactan significativamente en la ocurrencia de este tipo de accidentes, por lo que se recomienda seguir profundizando en el análisis temporal y territorial para futuras campañas de prevención.

Predicción de accidentes por zona urbana (Modelo2)

Algoritmos utilizados: Regresión Lineal y Random Forest Regressor.

Variables empleadas: Año, mes, total de accidentes, alcohol positivo, cantidad de lesionados.

Ajuste de hiper parámetros: Valores estándar.

Evaluación del Modelo a implementar

Se dividió el dataset en entrenamiento (80%) y testeo (20%). Las métricas utilizadas fueron:

- MAE (Error Absoluto Medio)
- RMSE (Raíz del Error Cuadrático Medio)
- R^2 (Coeficiente de determinación)

***Modelo 2 Predicción de Accidentes en Zona Urbana ***

```
features_2 = ['Año', 'Mes_num', 'Total', 'Alcohol_positivo', 'Cant_lesionados']
X2 = df[features_2]
y2 = df['Zona_urbana']

X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, test_size=0.2, random_state=42)

lr2 = LinearRegression()
lr2.fit(X2_train, y2_train)
y2_pred_lr = lr2.predict(X2_test)

rf2 = RandomForestRegressor(random_state=42)
rf2.fit(X2_train, y2_train)
y2_pred_rf = rf2.predict(X2_test)

print_metrics(y2_test, y2_pred_lr, "Regresión Lineal (Zona Urbana)")
print_metrics(y2_test, y2_pred_rf, "Random Forest (Zona Urbana)")
```

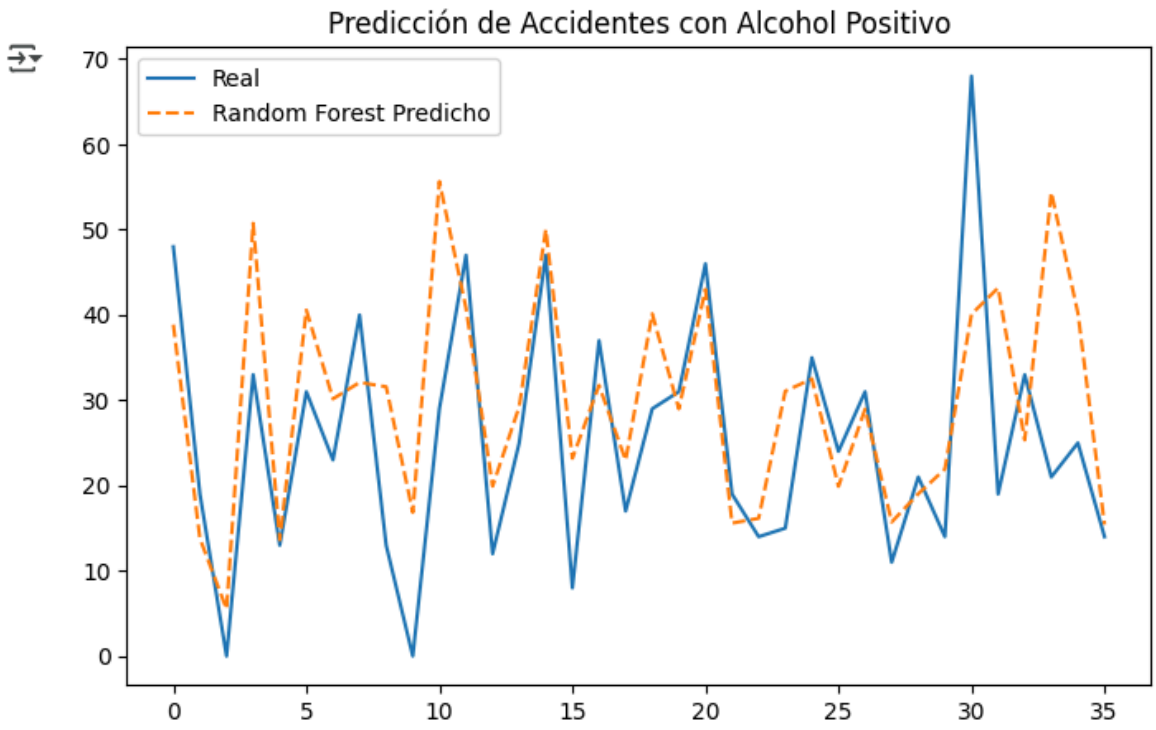
```
--- Métricas Regresión Lineal (Zona Urbana) ---
MAE: 6.112502181660647
RMSE: 10.126299087261499
R²: 0.9934266232574552

--- Métricas Random Forest (Zona Urbana) ---
MAE: 4.309444444444444
RMSE: 6.748185764419286
R²: 0.9970808186096701
```

En el segundo modelo las métricas de regresión lineal para ambas zonas muestran que este modelo explica aproximadamente el 99.3% de la variabilidad en la cantidad de accidentes en zona urbana. Tanto el error absoluto medio como el cuadrático son bajos, lo que indica un excelente ajuste, el modelo Random Forest mejora aún más los resultados, explicando el 99.8% de la variabilidad de la variable objetivo y reduciendo tanto el MAE como el RMSE.

Resultados Reales vs predichos para alcohol positivo

```
[ ]  plt.figure(figsize=(8,5))
     plt.plot(y1_test.values, label='Real')
     plt.plot(y1_pred_rf, label='Random Forest Predicho', linestyle='dashed')
     plt.title('Predicción de Accidentes con Alcohol Positivo')
     plt.legend()
     plt.show()
```



El gráfico compara los valores reales de accidentes con alcohol positivo con los valores predichos por el modelo Random Forest sobre el conjunto de datos.El modelo Random Forest logra seguir la tendencia general de los datos reales, anticipando correctamente muchos de los picos y valles. En ciertas partes del gráfico hay momentos en los que las predicciones difieren de los valores reales, mostrando la dificultad de anticipar con precisión los accidentes relacionados al alcohol.

El modelo Random Forest presenta un desempeño razonablemente bueno para predecir la tendencia y magnitud de los accidentes con alcohol positivo, aunque su precisión podría mejorar si se incorporan variables adicionales como eventos especiales, controles policiales, clima o campañas de prevención.

Conclusión Final :

El modelo permite anticipar tendencias generales de accidentes y casos con alcohol positivo, apoyando la toma de decisiones en políticas públicas,la baja correlación entre alcohol positivo y otras variables limita la capacidad predictiva, sugiriendo que se requieren datos adicionales (horarios, condiciones, clima, eventos) para mejorar el modelo.
La importancia de variables en Random Forest destaca la relevancia del total de accidentes, la zona y el mes para anticipar eventos críticos.
El uso de modelos predictivos debe institucionalizarse y actualizarse periódicamente para mejorar la gestión de la seguridad vial.

Recomendaciones :

Aumentar controles en zona urbana , realizar campañas de concientización sobre el consumo de alcohol en la conducción sobre todo en los meses donde se observa el aumento de casos.

mejorar la recoleccion de datos sobre todo en las variables faltantes como sexo,edad, clima, estado de la calzada,ubicacion de la colision,hora,conductor posee Licencia Nacional de Conducir, si posee la categoria correspondiente, falta cometida.

Los mencionados anteriormente son indispensables para enriquecer la fuente de datos para que al momento de utilizar técnicas de aprendizaje automático se pueda lograr mejores resultados en la predicción y en la toma de decisiones informadas.