Proyecto de Ciencia de Datos: Análisis de Accidentes de Tránsito en Bucaramanga

Oscar Augusto Diaz Triana

Septiembre 2023

Introducción

- Buenas tardes a todos, estoy emocionado de presentarles nuestro proyecto de ciencia de datos que se centra en la seguridad vial en la ciudad de Bucaramanga, Colombia. En un mundo donde la movilidad es esencial, la seguridad en las carreteras es de suma importancia. Nuestro proyecto aborda este desafío al analizar detenidamente los accidentes de tránsito en Bucaramanga y proponer soluciones basadas en datos para mejorar la seguridad vial en la región.

Descripción del Proyecto

- Para comprender la magnitud de nuestro proyecto, es crucial entender la problemática que enfrentamos. Bucaramanga, como muchas otras ciudades, enfrenta desafíos relacionados con la seguridad en las carreteras. Los accidentes de tránsito son un problema persistente que afecta a la comunidad, y es aquí donde entra en juego nuestro proyecto. Nos enfocamos en analizar estos accidentes para identificar patrones, factores de riesgo y tendencias, lo que nos permitirá desarrollar recomendaciones efectivas para mejorar la seguridad vial en la ciudad.

Datos Utilizados

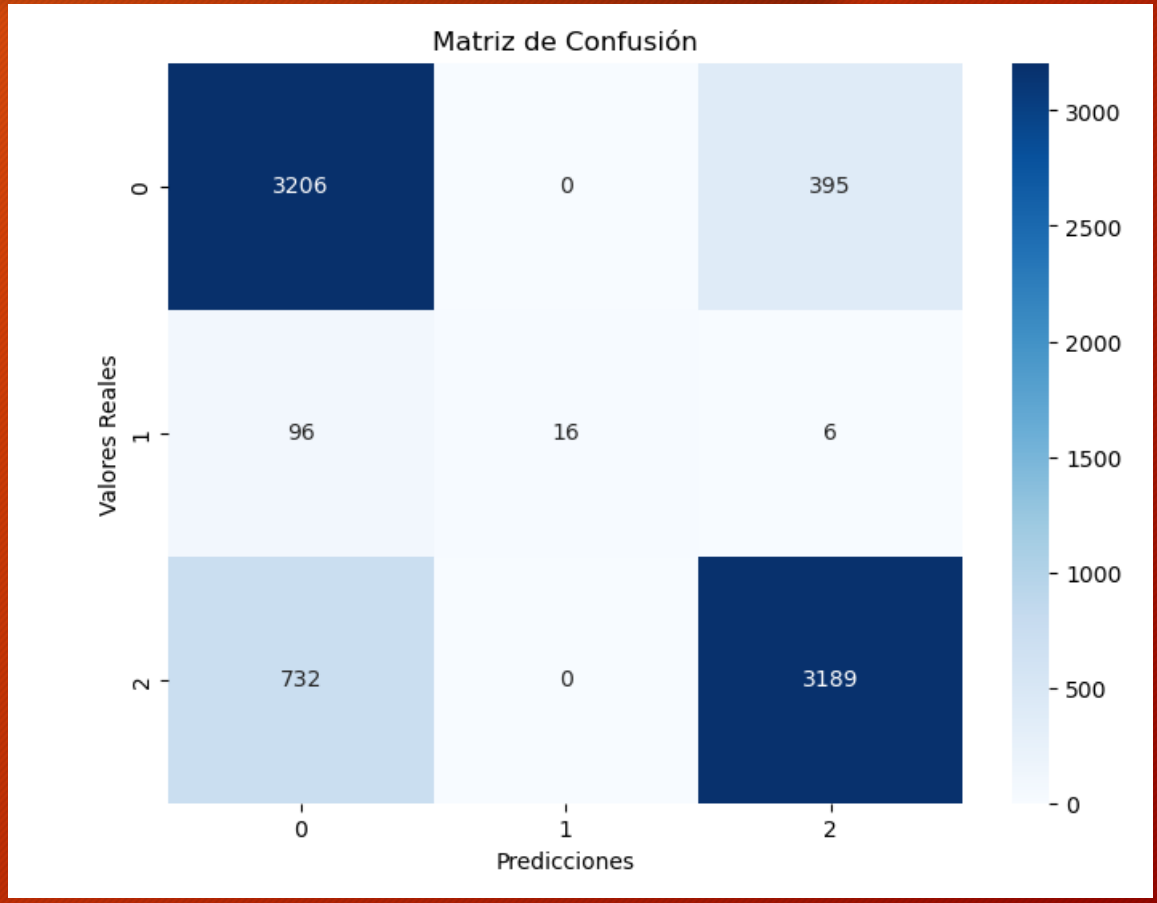
- Los cimientos de nuestro proyecto descansan en datos detallados de accidentes de tránsito en Bucaramanga. Obtuvimos estos datos de fuentes confiables, incluida la plataforma del gobierno colombiano, Datos Abiertos. Este conjunto de datos es rico en información, que incluye la ubicación geográfica, fecha y hora de los accidentes, condiciones climáticas, tipos de vehículos involucrados, y la causa probable de los accidentes, entre otros detalles clave.

Metodología

- Nuestra metodología se basó en un enfoque de ciencia de datos sólido. Comenzamos por cargar y preprocesar los datos para que sean aptos para el análisis. Luego, utilizamos técnicas de agrupamiento para identificar patrones geográficos y modelos de clasificación para predecir la gravedad de los accidentes. Estos procesos nos permitieron obtener conocimientos valiosos para abordar la seguridad vial en Bucaramanga de manera efectiva.

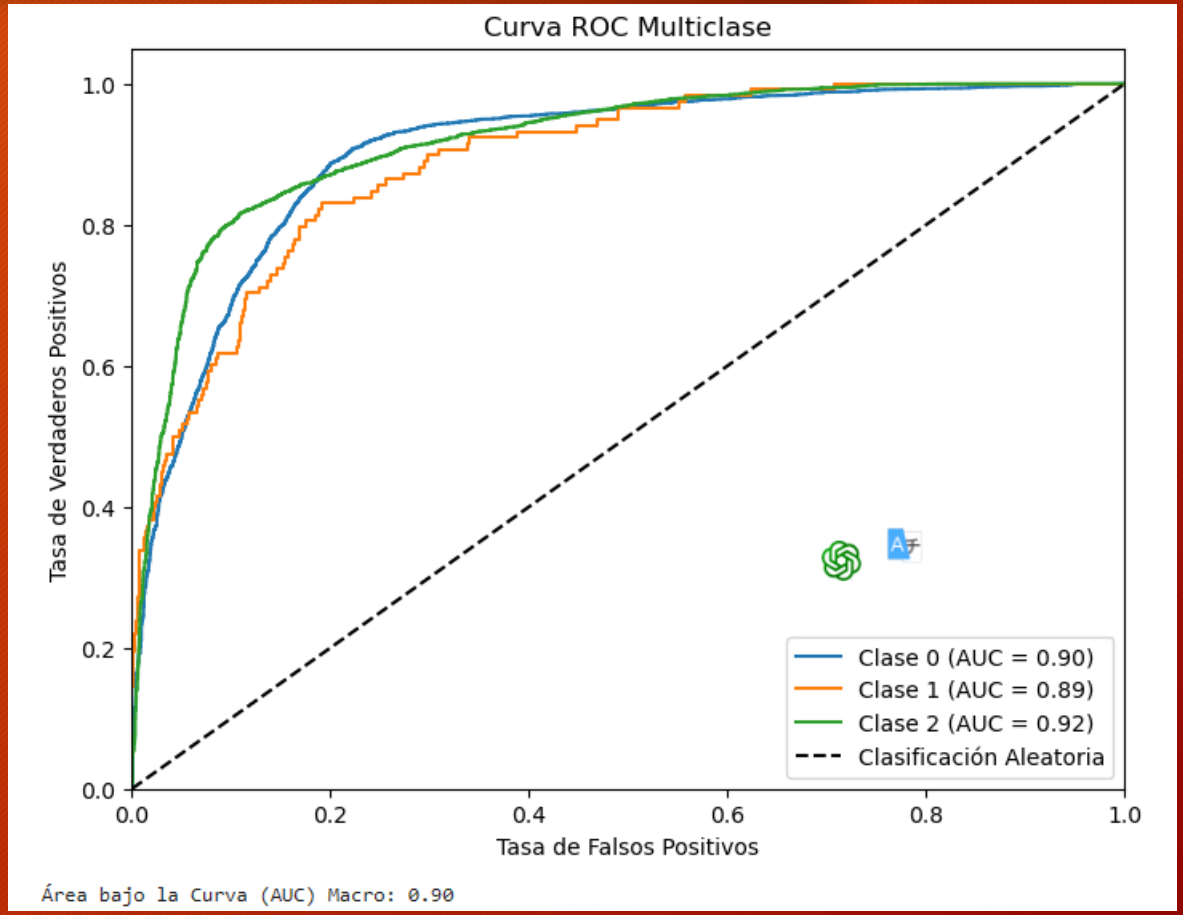
Modelo de Clasificación

- Inicialmente el modelo CatBoost muestra un buen rendimiento en la clasificación de las clases "Con Daños" y "Con Muertos", con altas precisiones y recalls. Sin embargo, tiene dificultades para identificar correctamente la clase "Con Heridos", lo que sugiere la necesidad de ajustes específicos para mejorar su rendimiento en esta clase.



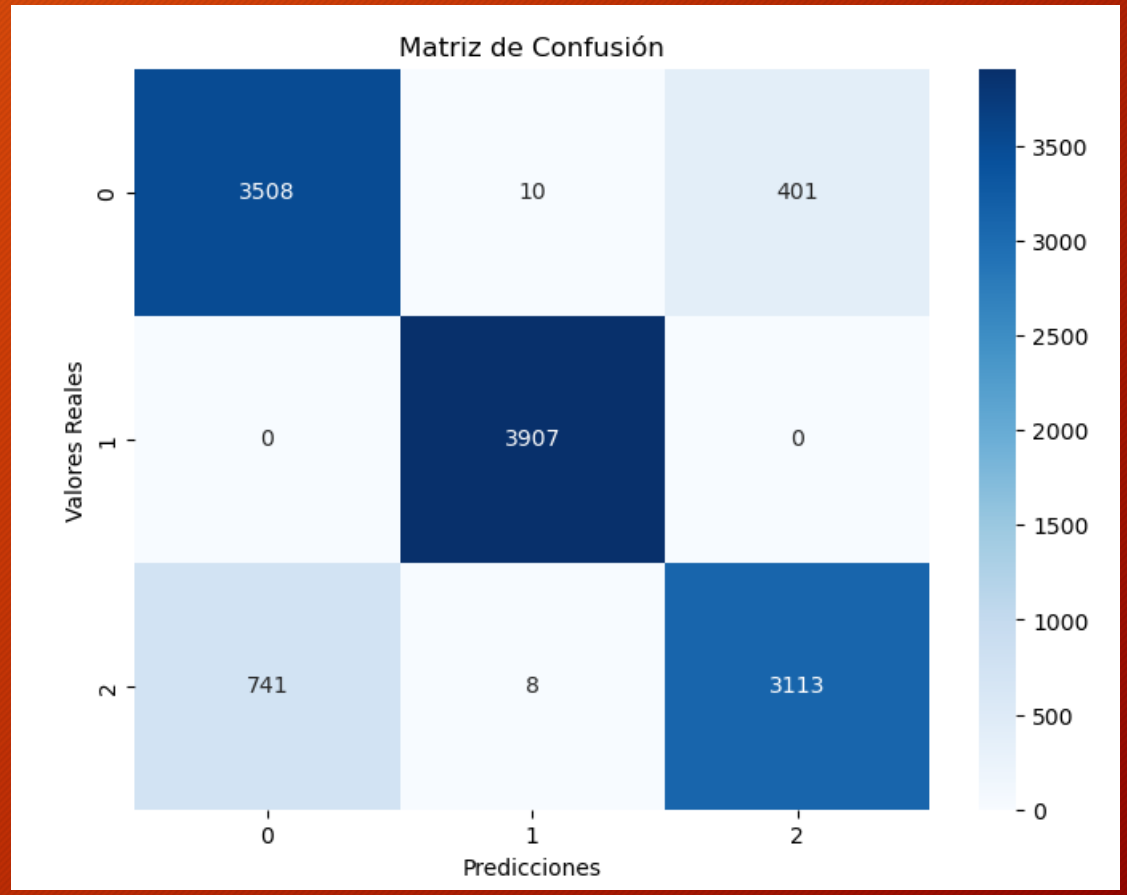
Curva ROC Multiclase

- Un AUC (Área bajo la Curva) de 0.90 es indicativo de un buen rendimiento del modelo en términos de su capacidad para distinguir entre las clases en el problema de clasificación multiclase. Las AUC individuales para cada clase también muestran que el modelo tiene una buena capacidad de discriminación para cada clase por separado.
- El AUC Macro de 0.90 sugiere que el modelo tiene un buen rendimiento general en todas las clases, lo cual es muy positivo. La curva ROC y los valores de AUC proporcionan información valiosa sobre la calidad del modelo y su capacidad para realizar clasificaciones precisas.



Manejo de Desequilibrio de Clases

- Después de abordar el desequilibrio de clases, el modelo mejoró significativamente su rendimiento. Ahora, muestra altas precisiones, recalls y F1-scores para todas las clases, lo que indica una clasificación precisa y equilibrada. El modelo alcanza una precisión general del 90%, lo que demuestra su eficacia en la predicción de las clases de seguridad vial.



Closs Validation Modelo catboost

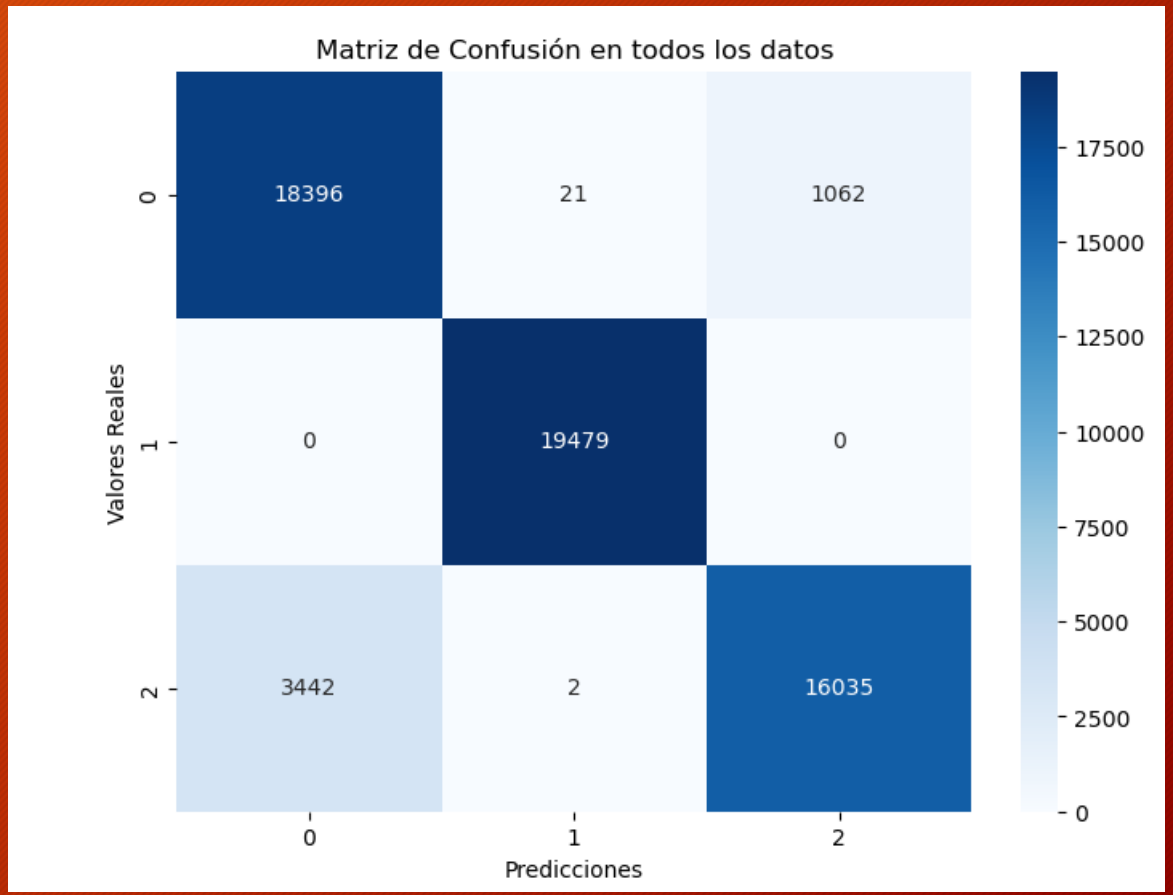
Se evaluó el rendimiento del modelo CatBoostClassifier en diversas configuraciones mediante cross-validation. Se analizaron los resultados de las siguientes configuraciones:

1. **500 iteraciones, 10 grupos de cross-validation:**
    1. Alto rendimiento en precisión, recall y F1-score en todas las clases.
    2. Precisión general del 92%.
    3. Promedio ponderado y no ponderado de métricas elevados.
2. **500 iteraciones, 15 grupos de cross-validation:**
    1. Resultados similares a la configuración anterior de 10 grupos.
    2. Rendimiento consistente en todas las clases.
3. **1000 iteraciones, 5 grupos de cross-validation:**
    1. Buen rendimiento en precisión y recall, pero ligera disminución en F1-score.
    2. Promedio ponderado y no ponderado aún altos, aunque ligeramente más bajos.
4. **1000 iteraciones, 10 grupos de cross-validation:**
    1. Resultados similares a la configuración de 5 grupos, pero con un rendimiento ligeramente mejor.
5. **1000 iteraciones, 15 grupos de cross-validation:**
    1. Resultados prácticamente idénticos a la configuración de 10 grupos.
6. **1500 iteraciones, 5/10/15 grupos de cross-validation:**
    1. Resultados consistentes con las configuraciones anteriores de 1000 iteraciones.

```python
from sklearn.model_selection import cross_val_score

# Realiza cross-validation y muestra los resultados
scores = cross_val_score(catboost_model, X_resampled, y_resampled, cv=5, scori
print("Accuracy en cada fold:", scores)
print("Accuracy promedio:", scores.mean())

# Entrena el modelo CatBoost en todos los datos de entrenamiento
catboost_model.fit(X_resampled, y_resampled)

# Realiza predicciones en el conjunto de prueba
y_pred = catboost_model.predict(X_resampled)

# Muestra el informe de clasificación en todos los datos
print("Informe de Clasificación en todos los datos:")
print(classification_report(y_resampled, y_pred))

# Matriz de Confusión en todos los datos
confusion = confusion_matrix(y_resampled, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(confusion, annot=True, fmt="d", cmap="Blues")
plt.title("Matriz de Confusión en todos los datos")
plt.xlabel("Predicciones")
plt.ylabel("Valores Reales")
plt.show()
```

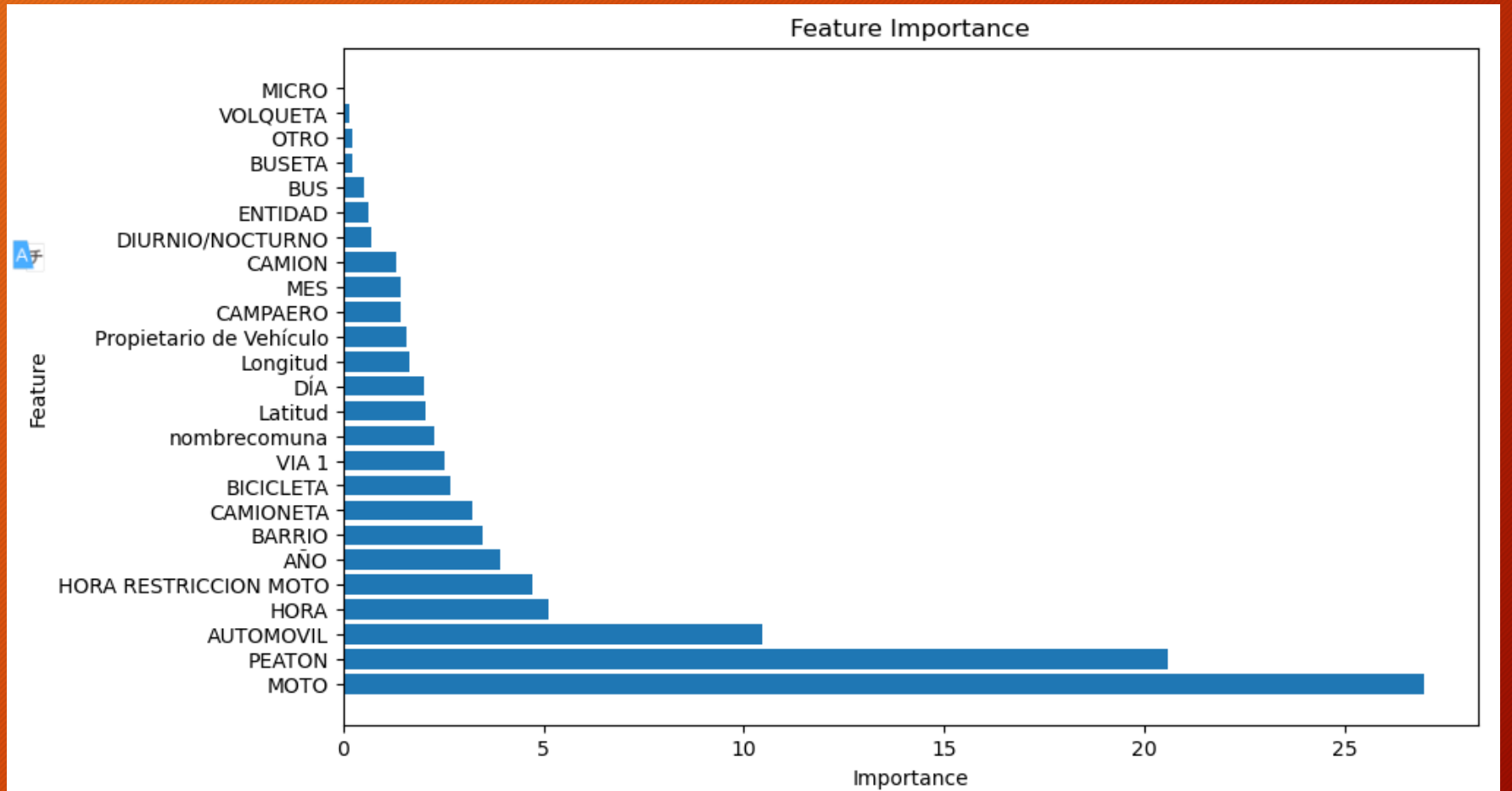
Mejores parámetros Closs Validation Modelo catboost

- La combinación que produjo los mejores resultados en precisión, recall y F1-score fue la siguiente:
- **Configuración: 500 iteraciones, 10 grupos de cross-validation:**
  - Precisión del 92% en todas las clases.
  - Alto rendimiento en recall y F1-score.
  - Promedio ponderado y no ponderado de métricas elevados.
- Aunque otras configuraciones mostraron resultados competitivos, esta configuración en particular parece ser la más equilibrada en términos de rendimiento y tiempo de entrenamiento. La elección final dependerá de las necesidades y recursos disponibles.

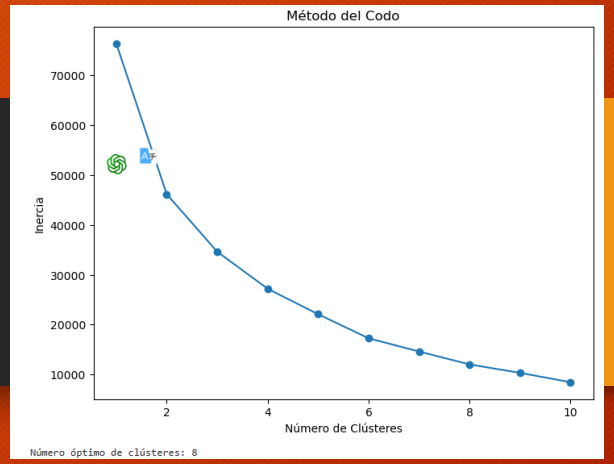


Visualización de Importancia de Características

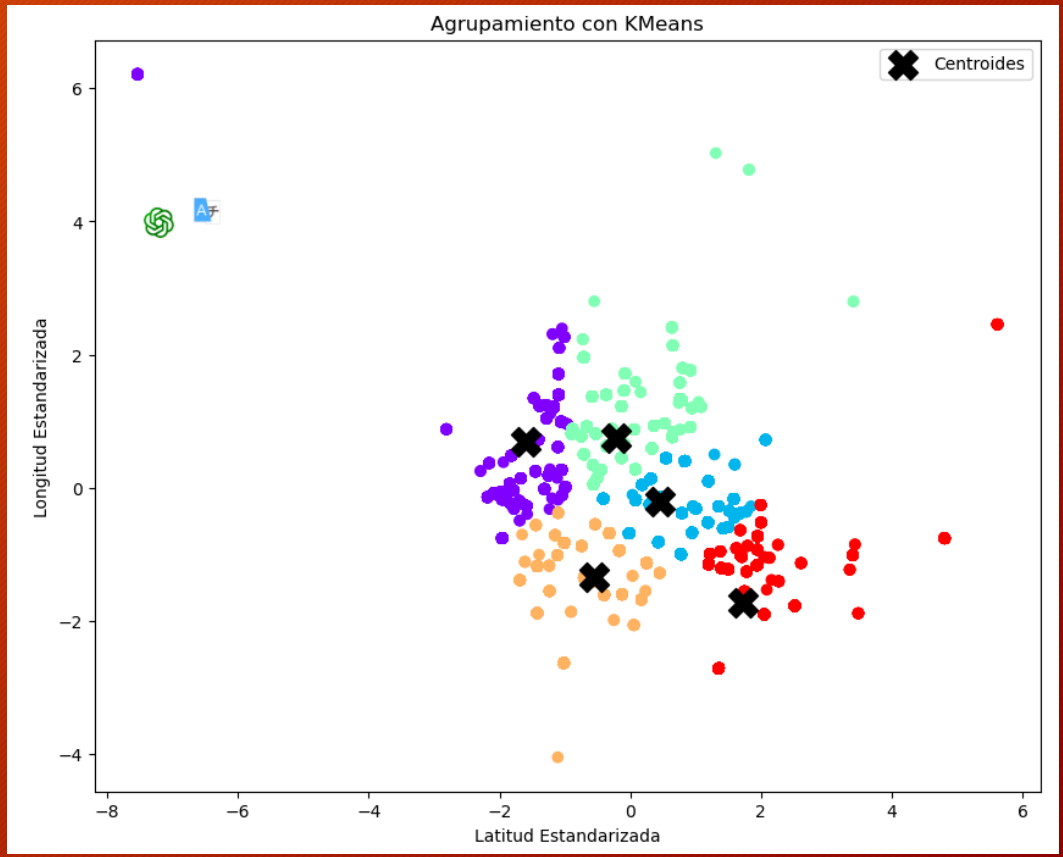
- Después de ajustar el modelo, visualiza la importancia de las características para comprender cuáles están influyendo más en las predicciones del modelo. Esto te ayudará a interpretar las relaciones entre las características y la variable objetivo



Creación y Evaluación de un Modelo para agrupamiento K-Means

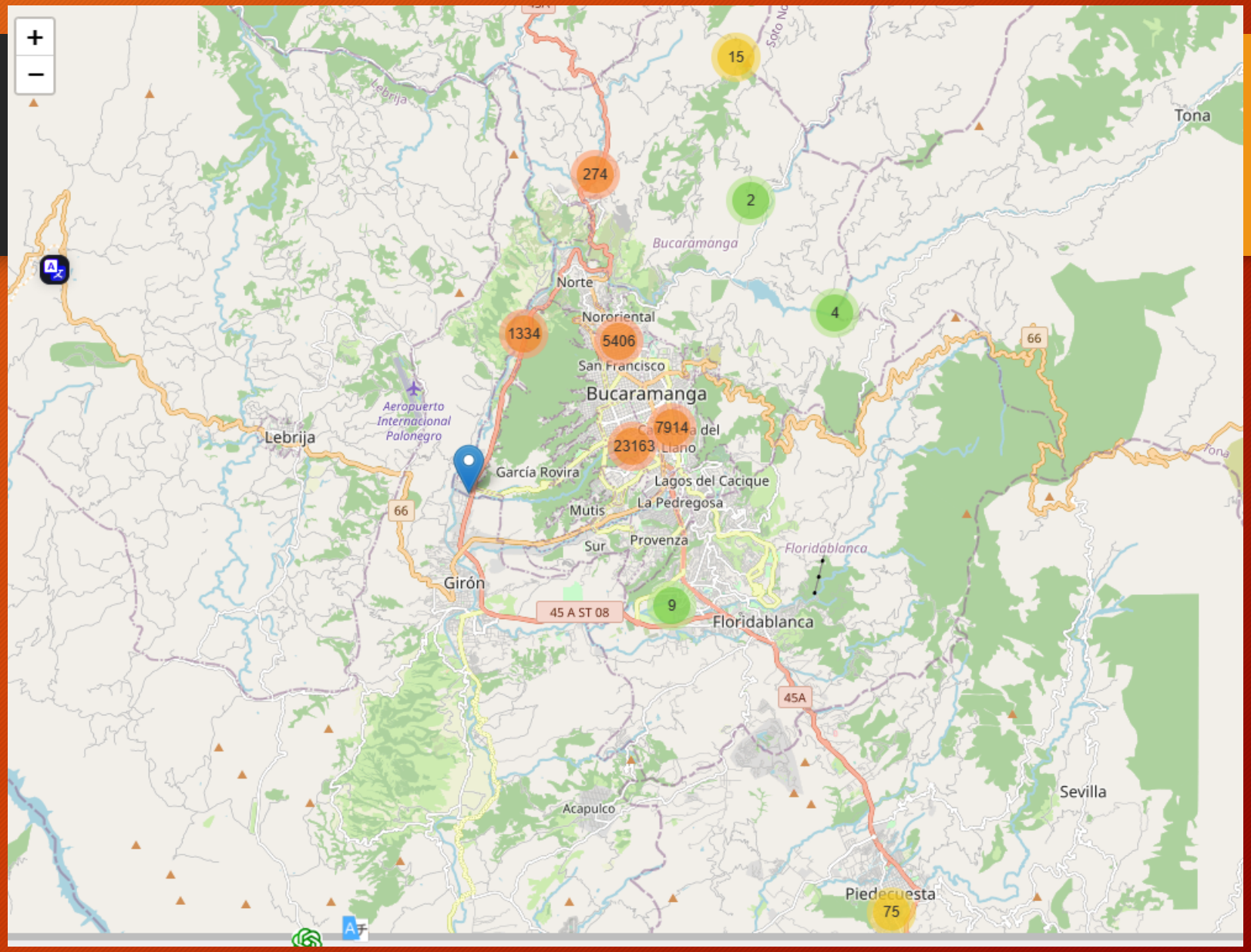


se elige un número óptimo de clústeres, que puede variar según lo que se observe en el gráfico generado por el método del codo. Luego, se aplica KMeans con el número seleccionado de clústeres a los datos escalados. Se muestran los resultados del agrupamiento, incluyendo el número de muestras en cada clúster. Finalmente, se visualizan los resultados en un diagrama de dispersión donde cada punto representa una muestra, y se muestran los centroides de los clústeres en forma de marcadores 'X'. Esto permite comprender la distribución de las muestras en los clústeres en un espacio bidimensional de latitud y longitud estandarizada.



Visualizaciones y Mapas

- Grupos de accidentes en mapa



Resultados

Agrupamiento de Barrios con KMeans

- Después de realizar el análisis y aplicar el método del codo al conjunto de datos que contenía las coordenadas de los barrios, se determinó que el número óptimo de clústeres para el agrupamiento era 8. Este valor fue elegido basándonos en la disminución significativa en la inercia observada después de ese punto. Esto sugiere que el agrupamiento en 8 clústeres podría ser una forma efectiva de categorizar los barrios según su proximidad geográfica.
- Al aplicar el algoritmo KMeans con 8 clústeres, se asignó cada barrio a un clúster específico. Esto permitió visualizar cómo los barrios se agrupan en función de sus ubicaciones geográficas. Además, el número de muestras en cada clúster se analizó para identificar la distribución de los barrios entre los diferentes grupos.

Modelado de Clasificación con CatBoost

- Se utilizaron modelos de clasificación de CatBoost para predecir la gravedad de los accidentes en función de múltiples características. Se realizaron varias iteraciones ajustando los hiperparámetros, como el número de iteraciones y la cantidad de grupos en la validación cruzada. Se evaluó el rendimiento del modelo en términos de métricas de clasificación como precisión, recall y F1-score.
- Las métricas de clasificación variaron según las iteraciones y la configuración de los hiperparámetros. Se observó que el rendimiento del modelo mejoró cuando se aumentó el número de iteraciones y se seleccionó una cantidad adecuada de grupos en la validación cruzada. Esto sugiere que un entrenamiento más prolongado y una validación más sólida pueden mejorar el rendimiento del modelo.

Comparación de Resultados

- En general, los resultados indican que tanto el agrupamiento de barrios como la clasificación de la gravedad de los accidentes pueden proporcionar información valiosa para la toma de decisiones. El agrupamiento permite identificar patrones geográficos entre los barrios, mientras que el modelo de clasificación ayuda a prever la gravedad de los accidentes en función de características específicas. Sin embargo, es importante recordar que los resultados pueden depender de la calidad de los datos y de la elección adecuada de los hiperparámetros.

Conclusiones y Recomendaciones

Agrupamiento de Barrios con KMeans:

- Mediante el método del codo, determinamos que el número óptimo de clústeres para agrupar los barrios era 8, lo que refleja patrones geográficos efectivos.
- El uso de KMeans nos permitió visualizar cómo se agrupan los barrios en función de su ubicación geográfica, lo que puede ayudar en la identificación de áreas problemáticas.

Modelado de Clasificación con CatBoost:

- Empleamos modelos de clasificación CatBoost para predecir la gravedad de los accidentes. La evaluación de múltiples configuraciones de hiperparámetros destacó la importancia de un mayor número de iteraciones y una sólida validación cruzada para mejorar el rendimiento.

Comparación de Resultados:

- Los resultados resaltan la utilidad tanto del agrupamiento de barrios como de la clasificación de gravedad de accidentes en la toma de decisiones relacionadas con la seguridad vial.
- Sin embargo, es crucial tener en cuenta la calidad de los datos y seleccionar adecuadamente los hiperparámetros para obtener resultados confiables.

Recomendaciones Futuras:

- Continuar mejorando la calidad de los datos utilizados en el proyecto, ya que esto podría conducir a resultados aún más precisos.
- Explorar otras técnicas de agrupamiento y modelado predictivo para ampliar la perspectiva y evaluar su eficacia en la seguridad vial.
- Considerar la implementación de medidas y estrategias basadas en los resultados obtenidos para contribuir a la prevención y mitigación de accidentes de tráfico en Bucaramanga.