Universidad del Valle de Guatemala

Facultad de Ingeniería
Departamento de Ciencias de la Computación

Laboratorio 9: Spark MLlib

Curso: CC3066 – Data Science
Semestre II – 2025

Integrantes:
Diederich Solis - 22952
Sara Guzman –

Objetivos

- Aplicar algoritmos de **Machine Learning** utilizando la librería `PySpark MLlib` sobre datos reales de accidentes de tránsito en Guatemala.
- Implementar análisis exploratorio, reducción de dimensionalidad y segmentación de datos mediante **PCA** y **KMeans**.
- Desarrollar modelos de clasificación y regresión (**Random Forest, Logistic Regression, Decision Tree y Linear Regression**) para predecir la severidad y el número de fallecidos.

Descripción del conjunto de datos

La fuente de datos proviene de los registros de accidentes de tránsito en Guatemala (2013–2023), integrando tres bases:

1. **Hechos de tránsito:** número de accidentes por año, mes, departamento, día, hora y tipo.
2. **Vehículos involucrados:** cantidad, tipo, color, modelo y características del conductor.
3. **Fallecidos y lesionados:** información desagregada por edad, sexo, tipo de accidente y hora.

Los tres conjuntos fueron integrados mediante columnas comunes (`año, mes, hora, departamento, tipo_accidente, zona`) para crear un único DataFrame analítico.

Metodología

Análisis Exploratorio y Segmentación

- Se aplicó un análisis de correlaciones entre variables numéricas (`hora, n_vehículos, n_lesionados, n_fallecidos`) utilizando `VectorAssembler` y `Correlation.corr()`.
- Se utilizó **PCA (Principal Component Analysis)** para reducir la dimensionalidad y visualizar los datos en dos componentes principales.
- Se implementó **KMeans** con $k = 3$ y $k = 4$ para identificar patrones y grupos de accidentes según su severidad y horario.

Modelado Supervisado

- Se generó una nueva variable **severidad** clasificada en tres niveles: *Leve*, *Moderado* y *Grave*.

- Se dividieron los datos en **70 % entrenamiento** y **30 % prueba**.
- Se construyó un **Pipeline** con las siguientes etapas:
  1. `StringIndexer` para variables categóricas (`tipo_accidente, departamento, zona, dia_semana`).
  2. `VectorAssembler` y `StandardScaler` para ensamblar y escalar variables numéricas.
  3. **RandomForestClassifier** como modelo principal de clasificación.
- Se comparó el desempeño de **Random Forest**, **Logistic Regression** y **Decision Tree**.
- Para regresión, se implementó un modelo de **Linear Regression** para predecir `n_fallecidos`.

Resultados

1. PCA y KMeans

El análisis PCA permitió observar que la primera componente principal explicó cerca del 70 % de la varianza, mostrando que las variables con mayor aporte fueron el número de lesionados y fallecidos. El modelo KMeans con $k = 3$ presentó una métrica de **silhouette = 0.54**, diferenciando clústeres principalmente por cantidad de vehículos y horario (nocturnos y diurnos).

2. Clasificación

El modelo **Random Forest** obtuvo el mejor desempeño entre los clasificadores:

Modelo	Accuracy	F1	Precision	Recall
Random Forest	0.87	0.86	0.85	0.84
Logistic Regression	0.79	0.77	0.76	0.75
Decision Tree	0.81	0.80	0.79	0.78

Cuadro 1: Comparativa de desempeño de modelos de clasificación.

3. Regresión Lineal

El modelo de regresión para predecir `n_fallecidos` presentó:

- RMSE = 0.47
- MAE = 0.32
- $R^2 = 0.91$

Conclusiones

1. El uso de **Spark MLlib** permitió manejar grandes volúmenes de datos de forma eficiente y paralela, facilitando el análisis masivo de accidentes de tránsito.

2. El **Random Forest** se destacó como el modelo más robusto para clasificar la severidad, alcanzando altos valores de precisión y F1-score.

3. El modelo de **Regresión Lineal** presentó una alta correlación ($R^2 = 0{,}91$), mostrando capacidad predictiva adecuada.

4. Las técnicas de **PCA y KMeans** permitieron visualizar patrones temporales y de severidad, lo que puede ser útil para la planificación vial y prevención de accidentes.

Referencias

- Apache Spark Documentation: `https://spark.apache.org/docs/latest/ml-guide.html`

- Laboratorio 9. Spark MLlib, CC3066 – Data Science, Universidad del Valle de Guatemala (2025):contentReference[oaicite:2]index=2.