	Colegio Universitario de Cartago	
	BIG DATA (BD)	

II Cuatrimestre 2025
Programación II
BD-143
Profesor Osvaldo González Chaves
Proyecto III

Objetivos del proyecto

Diseñar y desarrollar proyectos de ciencia de datos en Python enfocados en problemáticas relevantes para Costa Rica, como el análisis accidentes tránsito, turismo, cambio climático, transporte público, electricidad, entre otros, integrando diversas fuentes de datos como bases de datos relacionales (SQLite, SQL Server, Postgresql o Mysql), APIs públicas costarricenses e internacionales, y archivos csv reales. Los proyectos incluirán un análisis exploratorio de datos (EDA), visualización de datos y la aplicación de algoritmos de machine learning supervisado. Todo el desarrollo se estructurará utilizando principios de programación orientada a objetos para fomentar buenas prácticas de diseño y mantenimiento del código.

Rubros de calificación

Criterio	**Porcentaje**
Uso de datos (CSV, APIs públicas, limpieza de datos)	25%
Conexión e integración con base de datos (SQLite, SQLServer, etc.)	25%
Análisis exploratorio de datos (EDA)	15%
Visualización de datos (gráficos, tendencias, comparaciones)	10%
Implementación de modelo de Machine Learning supervisado	15%
Dashboard interactivo con Streamlit(gradio , dash o NiceGUI)	*10%*
+*Uso de mapas como visualización adicional*	*2%*

Total: 100%

Proyecto 1: Optimización y Predicción de Demanda del Transporte Público

Descripción: Analizar patrones de uso del transporte público para predecir cantidad de pasajeros y optimizar frecuencias de servicio.

Fuentes de Datos:

- **CSV:** Estadísticas INCOFER: https://datos.incofer.go.cr/dataset/estadisticas-pasajeros
- **API:** Feriados CR: https://api-feriados-cr.herokuapp.com/api/2025
- **Base de datos:** PostgreSQL con validaciones por estación y ocupación histórica.

EDA y Visualización:

- Análisis de flujo de pasajeros por hora pico y valle
- Identificación de rutas saturadas
- Visualización de demanda por día de la semana
- Impacto de lluvia en cantidad de pasajeros

Modelo Supervisado - Regresión:

- **Tipo**: Regresión (predicción del número de pasajeros)
- **Algoritmos**: Regresión lineal, KNN, Random Forest, XGBoost
- **Variables de entrada**: Hora del servicio, estación, día tipo (laboral/feriado), precipitación, temperatura
- **Variable objetivo**: Número de pasajeros por servicio

Modelo Supervisado - Clasificación:

- **Tipo:** Clasificación (nivel de ocupación del servicio)
- **Algoritmos:** Regresión logística, Árbol de decisión, Random Forest
- **Variables de entrada:** Número de pasajeros, capacidad del vehículo, hora, ruta
- **Variable objetivo:** Nivel de ocupación (Baja/Media/Alta/Saturada)

Proyecto 2: Análisis y Predicción del Turismo en Costa Rica

Descripción: Incluye análisis exploratorio y visualizaciones previas al modelado. Este proyecto busca analizar el comportamiento del turismo en Costa Rica y predecir la cantidad de visitantes anual o mensual, considerando factores como el clima, el país de origen de los turistas y eventos relevantes.

Fuentes de Datos:

- CSV: Instituto Costarricense de Turismo: https://www.ict.go.cr/es/estadisticas/informes-estadisticos.html
- API de clima: Open-Meteo: https://api.open-meteo.com/v1/forecast?latitude=9.93&longitude=-84.08&daily=temperature_2m_max&timezone=America%2FCosta_Rica
- Base de datos: Turistas por mes, país y clima diario (almacenado en SQLite, SQL Server etc.).

EDA y Visualización:

- Explorar la distribución de turistas por mes y país.
- Visualizar tendencias de llegada según el clima.
- Mapas interactivos para mostrar regiones de origen y destino.

Modelo Supervisado:

- **Tipo**: Regresión (predicción del número de turistas).
- **Algoritmos**: Regresión lineal, KNN, Random Forest.
- **Variables de entrada:** Mes, país de origen, temperatura promedio, eventos.
- **Variable objetivo:** Número de turistas mensuales.

Proyecto 3: Análisis y Predicción de la Calidad del Aire en el GAM

Descripción: Analizar la calidad del aire en el Gran Área Metropolitana correlacionándola con datos de tráfico vehicular y condiciones meteorológicas para predecir niveles de contaminación.

Fuentes de Datos:

- CSV: Flujo vehicular MOPT: https://datos.mopt.go.cr:8080/dataset/flujo-vehicular
- Estadísticas COSEVI: https://datosabiertos.csv.go.cr/dashboards/19706/estadisticas-de-transito/
- APIs: Air Quality Open-Meteo: https://air-quality-api.open-meteo.com/v1/air-quality?latitude=9.9281&longitude=-84.0907&hourly=pm10,pm2_5,carbon_monoxide,nitrogen_dioxide,ozone
  Datos históricos clima: https://archive-api.open-meteo.com/v1/archive?latitude=9.9281&longitude=-84.0907&start_date=2024-01-01&end_date=2024-12-31
- Base de datos: SQLite con mediciones horarias de PM2.5, PM10, NO2, flujo vehicular, temperatura, humedad, velocidad del viento.

EDA y Visualización:

- Correlación entre flujo vehicular y niveles de PM2.5/NO2
- Análisis temporal de contaminación (patrones diarios/semanales)
- Mapas de calor de zonas críticas
- Gráficos de dispersión de variables meteorológicas vs contaminantes

Modelos Supervisados:

- **Tipo**: Regresión (predicción de concentración de PM2.5) ó Clasificación (categorización de calidad del aire)
- **Algoritmos**: Regresión lineal, Regresión múltiple, KNN, Random Forest, Regresión logística, árbol de decisión.
- **Variables de entrada**: Hora del día, flujo vehicular, temperatura, humedad, velocidad del viento, día de la semana ó Concentración PM2.5, NO2, O3, hora, temperatura, flujo vehicular
- **Variable objetivo**: Concentración de PM2.5 en µg/m³ ó Categoría ICA (Buena/Moderada/Mala/Muy Mala)

Proyecto 4: Predicción de Accidentes de Tránsito en Costa Rica

Descripción: Incluye análisis exploratorio y visualizaciones de zonas y condiciones. Analizar los factores que influyen en los accidentes de tránsito y predecir su ocurrencia por provincia, tipo de vía o condiciones climáticas.

Fuentes de Datos:

- CSV: https://www.csv.go.cr/estad%C3%ADsticas
- API climática: https://archive-api.open-meteo.com/v1/archive?latitude=9.93&longitude=-84.08&start_date=2023-01-01&end_date=2023-12-31&daily=precipitation_sum&timezone=America%2FCosta_Rica
- Base de datos: Accidentes por provincia, tipo, condiciones del clima (almacenado en SQLite, SQL Server etc.).

EDA y Visualización:

- Análisis de frecuencia de accidentes por provincia y hora del día.
- Visualización de mapas de calor por zona.
- Comparación entre tipos de accidentes y condiciones climáticas.

Modelo Supervisado:

- Tipo: Clasificación (si ocurrirá un accidente en cierta zona).
- Algoritmos: Regresión logística, árbol de decisión, KNN.
- **Variables de entrada:** Provincia, tipo de vía, lluvia acumulada, día de la semana, hora.
- **Variable objetivo:** Ocurrencia de accidente (Sí/No).

Proyecto 5: Predicción del Consumo de Energía en Costa Rica

Descripción: Incluye EDA con patrones horarios y visualización del impacto climático. Analizar el consumo de energía eléctrica por hora y predecir la demanda futura según variables como el clima, día de la semana, hora y tipo de zona.

Fuentes de Datos:

- CSV: https://aresep.go.cr/datos-abiertos/tarifas-electricidad-sistema-distribucion/

  https://aresep.go.cr/electricidad/tarifas/

- API climática: https://api.open-meteo.com/v1/forecast?latitude=10.0&longitude=-84.0&hourly=temperature_2m,cloudcover,windspeed_10m&timezone=America%2FCosta_Rica
- Base de datos: Datos de consumo horario y variables climáticas (almacenado en SQLite, SQL Server etc.).

EDA y Visualización:

- Análisis de consumo por hora y día de la semana.
- Gráficos de líneas para observar la demanda en distintas condiciones.
- Visualizaciones de correlaciones con el clima.

Modelo Supervisado:

- **Tipo**: Regresión (predicción del consumo en kWh).
- **Algoritmos**: Regresión lineal, KNN, XGBoost.
- **Variables de entrada:** Hora, día de la semana, temperatura, velocidad del viento, cobertura nubosa.
- **Variable objetivo:** Consumo energético (kWh).

Proyecto 6: Predicción de Deserción Estudiantil Universitaria

Descripción: Identificar estudiantes en riesgo de deserción y predecir rendimiento académico para intervenciones tempranas.

Fuentes de Datos:

- **CSV:** CONARE matrícula: https://datos.conare.ac.cr/dataset/matricula-educacion-superior
- **API:** INEC Pobreza: https://www.inec.cr/pobreza-y-desigualdad/pobreza-multidimensional
- **Base de datos:** PostgreSQL con historial académico.

EDA y Visualización:

- Tasas de deserción por carrera
- Identificación de materias filtro
- Correlación factores socioeconómicos-permanencia
- Dashboard de alertas tempranas

Modelo Supervisado - Clasificación:

- **Tipo:** Clasificación binaria (predicción de deserción)
- **Algoritmos:** Regresión logística, Árbol de decisión, Random Forest, XGBoost
- **Variables de entrada:** Nota admisión, promedio primer semestre, créditos aprobados/reprobados, tipo de beca, distancia universidad, índice pobreza cantonal
- **Variable objetivo:** Deserción (Sí/No)

Modelo Supervisado - Regresión:

- **Tipo**: Regresión (predicción del promedio ponderado)
- **Algoritmos**: Regresión lineal, Regresión múltiple, KNN, Random Forest
- **Variables de entrada**: Nota admisión, asistencia, créditos matriculados, tipo colegio, edad ingreso
- **Variable objetivo**: Promedio ponderado acumulado (0-100)

Estructura común de carpeta para todos los Proyectos (POO)

```
proyecto_nombre/
├── src/                    # Código fuente principal
│   ├── datos/              # Gestión de archivos y transformación de datos
│   ├── basedatos/          # Módulos para conexión a bases de datos
│   ├── api/                # Clientes para llamadas a APIs externas
│   ├── eda/                # Exploración de datos y estadísticas descriptivas
│   ├── visualizacion/      # Visualización de datos y mapas
│   ├── modelos/            # Entrenamiento y evaluación de modelos ML
│   ├── helpers/            # Funciones auxiliares reutilizables
│   └── main.py             # Punto de entrada del proyecto
│
├── notebooks/              # Jupyter notebooks para desarrollo y presentación
│   └── exploracion_inicial.ipynb
│
├── data/                   # Archivos (CSV, Excel, JSON, etc.)
    └── raw                 #Archivos en crudo
    └── processed           #Archivos procesados
```

Cada módulo debe implementarse usando clases que representen entidades o procesos:

- datos/ → Clase **GestorDatos**: encargada de cargar, transformar y exportar archivos CSV, Excel, etc.
- basedatos/ → Clase **GestorBaseDatos**: conecta con SQLite, MySQL, PostgreSQL o SQL Server y permite ejecutar consultas.
- api/ → Clase **ClienteAPI**: realiza peticiones a APIs públicas y transforma los resultados en DataFrames.
- eda/ → Clase **ProcesadorEDA**: realiza análisis estadístico, limpieza y exploración inicial de los datos.
- visualizacion / → Clase **Visualizador**: crea gráficos (líneas, barras, mapas, heatmaps, etc.).

- modelos/ → Clase **ModeloML**: entrena y evalúa modelos supervisados (regresión o clasificación).
- helpers/ → Clase **Utilidades**: contiene funciones auxiliares reutilizables para validaciones, formateo, etc.

Puntos extra:

- Creación de mapas como visualización adicional.