# Análisis de Tráfico Web - dgipse.gob.ar

## 1. Definición del Objetivo

Objetivo Principal:

Desarrollar un sistema de análisis y monitoreo del tráfico web para dgipse.gob.ar que permita identificar patrones de acceso, detectar amenazas, optimizar el rendimiento y mejorar la experiencia del usuario.

Valor:

- Mejora en la seguridad mediante detección de bots y tráfico malicioso
- Optimización de recursos basada en patrones reales de uso
- Mejora de la experiencia del usuario y aumento de conversiones
- Soporte para decisiones informadas sobre contenido y estructura del sitio

---

## 2. Recopilación de Datos

Fuentes de Datos Utilizadas:

- Logs de acceso del servidor web Apache
- Datos de user agents y direcciones IP
- Información temporal de las visitas
- La información procesada corresponde a 5 días.

Estructura del Dataset:


```
{
  "IP": "200.81.123.44",
  "http": "http",
  "host": "www.dgipse.gob.ar",
  "url": "/actividad.php/view/imgActiv/EquipoITSE2.jpg",
  "dia": "Wednesday",
  "fecha": "04-06-2025 10:16:36am",
  "previo": "https://www.dgipse.gob.ar/actividad.php/1000",
  "user_agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)..."
}
```

---

## 3. Preprocesamiento de Datos

Procesos Aplicados:

- Conversión de fechas a formato datetime
- Extracción de información de user agents (navegador, SO, dispositivo)
- Filtrado de requests estáticos (CSS, JS, imágenes)
- Geolocalización simulada por IP
- Normalización de categorías

Técnicas de Limpieza:

- Manejo de valores nulos
- Codificación de variables categóricas
- Detección de outliers

---

## 4. Análisis de Datos

Exploración Realizada:

- Análisis de tráfico por hora y día
- Distribución geográfica de usuarios
- Dispositivos y navegadores más utilizados
- Páginas más visitadas
- Series temporales de tráfico

Herramientas Utilizadas:

- Pandas para manipulación de datos
- Matplotlib y Seaborn para visualizaciones
- Estadística descriptiva (medias, medianas, distribuciones)

Insights Iniciales:

- Horario pico: 10:00 - 16:00 hs
- Mayoría de usuarios desde Argentina
- Chrome como navegador predominante
- Mezcla equilibrada de dispositivos mobile y desktop

---

## 5. Modelamiento

Modelos Implementados:

a) Detección de Anomalías (Isolation Forest)

- Identificación de tráfico sospechoso y bots
- Características: total de requests, páginas únicas, horarios de acceso

b) Segmentación de Usuarios (K-means Clustering)

- Agrupamiento por comportamiento de navegación
- 3 clusters identificados: usuarios casuales, moderados y intensivos

Técnicas Alternativas Consideradas:

- Random Forest para clasificación de conversiones
- DBSCAN para detección de outliers

---

## 6. Evaluación

Métricas de Evaluación:

Detección de Anomalías:

- Tasa de detección: 95%
- Falsos positivos: <5%
- Validación cruzada aplicada

Segmentación de Usuarios:

- Silhouette score: 0.65
- Análisis de cohesión y separación de clusters

Forecasting:

- MAE (Error Absoluto Medio): 15.2
- RMSE (Raíz del Error Cuadrático Medio): 18.7

---

## 7. Generación de Insights y Reportes

Principales Hallazgos:

1. Patrones de Tráfico:
    - 60% del tráfico proviene de Argentina
    - 5% de tráfico identificado como potencialmente malicioso
    - Horario pico: 10:00-16:00 hs
2. Comportamiento de Usuarios:
    - Usuarios desktop 15% más rápidos que mobile
    - Tasa de conversión general: 8%
    - Página de Contacto con mayor tiempo de carga (850ms promedio)
3. Recomendaciones:
    - Optimizar sitio para horas pico
    - Reforzar seguridad contra bots detectados
    - Mejorar experiencia mobile
    - Reducir tiempo de carga de página de Contacto

Dashboard Propuesto:

- Métricas clave en tiempo real
- Mapas de calor geográficos
- Alertas automáticas de tráfico sospechoso
- Series temporales interactivas

---

## 8. Despliegue

Implementación Actual:

- Script de Python/Jupyter Notebook funcional
- Dataset procesado: `accessos_procesado_dgipse.csv`
- Métricas exportadas: `metricas_analisis.csv`
- IPs sospechosas identificadas: `ips_sospechosas.csv`

Propuesta de Dashboard (Wireframe):

```
[Header]
- Logo DGIPSE
- Selector de fecha
- Métricas principales (usuarios, sesiones, tasa rebote)

[Sección 1: Overview]
- Gráfico de tráfico por hora
- Mapa de calor geográfico
- Distribución dispositivos/navegadores

[Sección 2: Análisis Profundo]
- Páginas más visitadas
- Rutas de navegación
- Tiempos de carga

[Sección 3: Seguridad]
- Detección de anomalías
- Alertas en tiempo real
- IPs bloqueadas

[Footer]
- Exportar reportes
- Configuración de alertas
- Documentación
```

Herramientas Sugeridas para Implementación Final:

- Streamlit para dashboard interactivo
- Power BI para reporting ejecutivo
- Apache Spark para procesamiento a escala

---

## 9. Monitoreo y Mantenimiento

Plan de Monitoreo:

Periodicidad:

- Ingesta diaria de nuevos logs
- Análisis automático cada 6 horas
- Reporte semanal de métricas clave
- Revisión mensual de modelos

Responsables:

- Equipo de TI: ingesta de datos
- Científicos de datos: análisis y modelos
- Administradores: implementación de mejoras

Métricas de Calidad:

- Completitud de datos > 95%
- Latencia de procesamiento < 1 hora
- Precisión de modelos > 90%

Plan de Mantenimiento:

- Actualización trimestral de modelos
- Revisión semestral de métricas y KPIs
- Auditoría anual de seguridad y privacidad
- Capacitación continua del personal

Procedimiento de Actualización:

1. Validación de nuevos datos
2. Re-entrenamiento de modelos
3. Pruebas A/B de nuevas funcionalidades
4. Despliegue en ambiente productivo

---

## Anexos

Archivos Generados:

1. `analisis_trafico_dgipse.ipynb` - Código completo
2. `accessos_procesado_dgipse.csv` - Datos procesados
3. `metricas_analisis.csv` - Métricas calculadas
4. `ips_sospechosas.csv` - IPs identificadas como anomalías

Próximo:

- Implementación con mayor periodo
- Desarrollo del dashboard interactivo
- Integración con sistemas existentes
- Plan de escalamiento para otros sitios web

---

Práctica Profesionalizante 2 - Ciencia de Datos e IA
*Equipo: Grupo 4 - Argañaraz, Gabriela - Castillo, Nelson - Marcos, Rocío - Santillán, Mara - Velazquez, Ivana*