PROYECTO DE INVESTIGACIÓN

Análisis de datos aplicado a problemáticas nacionales con R

EIY403 - Introducción al análisis de datos para otras carreras

Escuela de Informática y Computación

Información del proyecto
Curso: EIY403 - Introducción al análisis de datos para otras carreras
Valor: 25 % de la nota final
Modalidad: Equipos de 2 a 5 estudiantes
Fecha de entrega: Miércoles 5 de noviembre, 6:00 PM
Formato de entrega: R Markdown compilado a HTML + código fuente .Rmd
Envío: Correo institucional a `jordyab00@gmail.com`
Asunto: `ProyectoFinal_NombreEquipo_Tema`

1. Objetivos del proyecto

1.1. Objetivo general

Desarrollar un proyecto de investigación que aplique técnicas de análisis exploratorio de datos para estudiar una problemática nacional relevante, utilizando R como herramienta de análisis y generando insights significativos para la toma de decisiones.

1.2. Objetivos específicos

- Identificar y justificar una problemática nacional de interés público
- Obtener y procesar datasets reales de fuentes oficiales costarricenses
- Aplicar metodologías de análisis exploratorio de datos de forma sistemática
- Generar visualizaciones profesionales e interpretaciones contextualizadas
- Proponer recomendaciones fundamentadas en evidencia estadística
- Comunicar resultados de forma clara y profesional usando R Markdown

2. Selección temática

Áreas temáticas sugeridas

Seguridad vial y transporte:

- Accidentes de tránsito: factores, ubicaciones, tendencias temporales
- Análisis de siniestralidad por provincia, tipo de vehículo, condiciones
- Efectividad de medidas de seguridad vial implementadas

Economía y mercado laboral:

- Análisis salarial por sector, región, género, nivel educativo
- Empleo en zonas francas: evolución, impacto regional, sectores
- Brecha salarial y equidad en diferentes industrias
- Costo de vida y poder adquisitivo por provincia

Salud pública:

- Cobertura y acceso a servicios de salud por región
- Mortalidad y morbilidad: causas principales y tendencias
- Impacto de políticas de salud pública

Educación:

- Deserción estudiantil: factores y patrones regionales
- Acceso a educación superior y equidad educativa
- Infraestructura educativa y recursos por zona

Ambiente y sostenibilidad:

- Calidad del aire y contaminación atmosférica
- Gestión de residuos sólidos y reciclaje
- Indicadores de sostenibilidad ambiental

Desarrollo social:

- Pobreza multidimensional y desigualdad social
- Índices de desarrollo humano cantonal
- Acceso a servicios básicos: agua, electricidad, internet
- Migración interna y externa

Química industrial (opcional):

- Industria química costarricense: producción, exportaciones, empleo
- Análisis de la cadena de suministro de productos químicos
- Impacto ambiental de la industria química nacional
- Innovación y desarrollo tecnológico en el sector químico

3. Fuentes de datos obligatorias

Fuentes oficiales requeridas

Fuentes nacionales primarias:

- Instituto Nacional de Estadística y Censos (INEC)
- Ministerios públicos relacionados con el tema elegido
- Banco Central de Costa Rica (BCCR)
- Tribunal Supremo de Elecciones (TSE)
- Consejo de Seguridad Vial (COSEVI)
- Sistema de Información del Mercado Laboral (SOIMEL)

Fuentes complementarias aceptadas:

- Organismos internacionales (CEPAL, OMS, UNESCO)
- Datos gubernamentales de acceso abierto
- Estudios de universidades públicas costarricenses

4. Estructura y formato del documento

4.1. Especificaciones técnicas de formato

Formato obligatorio según IEEE

Configuración del documento R Markdown:

- **Fuente:** Times New Roman, 12 pt
- **Interlineado:** 1.5 líneas
- **Márgenes:** 2.5 cm en todos los lados
- **Sangría:** Primera línea de cada párrafo con sangría de 1.27 cm
- **Extensión total:** 8,000 - 12,000 palabras (aproximadamente 15-20 páginas)
- **Numeración:** Páginas numeradas en la esquina inferior derecha

Configuración YAML requerida para R Markdown:

```
output:
  html_document:
    toc: true
    toc_float: true
    toc_depth: 3
    number_sections: true
    theme: default
    highlight: tango
    code_folding: hide
    df_print: paged
    fig_caption: true
    css: "estilo_ieee.css"
```

4.2. Estructura obligatoria del documento

1. **Resumen ejecutivo** (200-300 palabras)
   - Problema investigado y justificación
   - Metodología aplicada
   - Principales hallazgos
   - Conclusiones y recomendaciones clave
2. **Introducción** (1,000-1,500 palabras)
   - Contextualización de la problemática nacional
   - Justificación de la relevancia del estudio
   - Objetivos específicos de la investigación
   - Preguntas de investigación planteadas
3. **Marco teórico y antecedentes** (1,500-2,000 palabras)
   - Conceptos fundamentales relacionados con el tema

- Estudios previos realizados en Costa Rica o región
- Marco metodológico del análisis exploratorio de datos
- Técnicas estadísticas aplicadas en el estudio

4. **Metodología** (1,000-1,500 palabras)
   - Descripción detallada de las fuentes de datos utilizadas
   - Proceso de obtención y preparación de los datos
   - Técnicas de limpieza y validación aplicadas
   - Análisis exploratorio sistemático implementado:
     a) Inspección inicial de los datos
     b) Evaluación de calidad y completitud
     c) Limpieza y preparación de variables
     d) Análisis univariado de variables clave
     e) Análisis bivariado y multivariado
     f) Síntesis de hallazgos y patrones identificados
   - Software utilizado (R, versión, librerías específicas)
5. **Análisis de resultados** (3,000-4,000 palabras)
   - Caracterización general del dataset
   - Análisis descriptivo de variables principales
   - Identificación de patrones, tendencias y outliers
   - Relaciones entre variables (correlaciones, asociaciones)
   - Análisis temporal cuando aplique
   - Análisis geográfico o regional cuando sea relevante
   - Interpretación contextualizada de todos los hallazgos
6. **Discusión de resultados** (1,500-2,000 palabras)
   - Interpretación de resultados en contexto nacional
   - Comparación con estudios previos y referencias internacionales
   - Limitaciones del estudio y los datos utilizados
   - Implicaciones de política pública
7. **Conclusiones y recomendaciones** (800-1,200 palabras)
   - Respuesta a las preguntas de investigación planteadas
   - Principales aportes del estudio
   - Recomendaciones específicas fundamentadas en evidencia
   - Propuestas para futuras investigaciones
8. **Referencias bibliográficas** (formato IEEE)
   - Mínimo 15 referencias, máximo 25
   - Al menos 8 fuentes oficiales costarricenses
   - Formato IEEE estrictamente aplicado
9. **Anexos**

- Código R completo y documentado
- Gráficos adicionales no incluidos en el texto principal
- Tablas complementarias de datos
- Diccionario de variables utilizado

5. Requisitos técnicos y de contenido

5.1. Requisitos del dataset

Características mínimas del dataset

- **Tamaño:** Mínimo 1,000 observaciones
- **Variables:** Al menos 8 variables de análisis
- **Diversidad:** Combinación de variables cuantitativas y cualitativas
- **Temporalidad:** Datos de al menos 2 años de período
- **Actualidad:** Datos no anteriores a 2020
- **Completitud:** Máximo 10 % de valores faltantes por variable
- **Relevancia:** Variables directamente relacionadas con la problemática

5.2. Requisitos de análisis técnico

- **Análisis descriptivo completo:** Medidas de tendencia central, dispersión y forma para todas las variables cuantitativas
- **Análisis de calidad:** Identificación y tratamiento de outliers, valores faltantes y inconsistencias
- **Visualizaciones profesionales:** Mínimo 10 gráficos de alta calidad (histogramas, boxplots, scatterplots, mapas cuando aplique)
- **Análisis bivariado:** Correlaciones, tablas de contingencia, análisis de varianza según corresponda
- **Interpretación contextual:** Cada resultado estadístico debe interpretarse en el contexto específico de la problemática
- **Código reproducible:** Todo el análisis debe ser completamente replicable con el código proporcionado

5.3. Requisitos de visualización

- Uso de paletas de colores profesionales y accesibles
- Títulos descriptivos y ejes claramente etiquetados
- Leyendas completas y comprensibles
- Tamaño de fuente legible (mínimo 10 pt)

- Uso apropiado del tipo de gráfico para cada variable
- Integración coherente de gráficos en el texto narrativo

6. Modalidad de equipos

Conformación de equipos

- **Tamaño:** 2 a 5 estudiantes por equipo
- **Registro:** Equipos deben registrarse antes del 15 de septiembre
- **Comunicación:** Designar un coordinador de equipo para comunicaciones oficiales
- **Responsabilidades:** Distribución equitativa de tareas entre todos los miembros
- **Individualidad:** Cada estudiante debe dominar completamente todo el proyecto

7. Proceso de consulta y seguimiento

Aunque el proyecto se califica en una sola entrega final, se ofrecen las siguientes oportunidades de consulta y seguimiento **opcionales**:

- **Consultas individuales:** Horarios de atención del profesor
- **Sesiones grupales:** Espacios de discusión en clase designados
- **Foro digital:** Preguntas técnicas vía correo electrónico
- **Revisión de avances:** Posibilidad de enviar borradores para retroalimentación (opcional)

Es responsabilidad de cada equipo aprovechar estos recursos según sus necesidades y gestión del tiempo.

8. Criterios de evaluación y rúbrica

Criterio	**Peso**	**Niveles de desempeño**			
		Excelente (90-100)	**Satisfactorio (70-89)**	**Aceptable (50-69)**	**Deficiente (0-49)**
Relevancia y justificación del problema	15 %	Problemática altamente relevante, justificación sólida con evidencia, impacto nacional claro	Problemática relevante, justificación adecuada, impacto identificado	Problemática moderadamente relevante, justificación básica	Problemática poco relevante, justificación débil o ausente

Calidad y procesamiento de datos	20 %	Dataset robusto de fuentes oficiales, limpieza impecable, validación rigurosa	Dataset apropiado, limpieza adecuada, validación presente	Dataset básico, limpieza mínima, poca validación	Dataset inadecuado, sin limpieza apropiada, sin validación
Metodología de análisis exploratorio	20 %	Análisis sistemático completo, técnicas apropiadas, interpretación experta	Análisis bien estructurado, técnicas correctas, buena interpretación	Análisis básico, técnicas limitadas, interpretación superficial	Análisis deficiente, técnicas incorrectas, interpretación pobre
Visualizaciones y gráficos	15 %	Visualizaciones profesionales, altamente informativas, diseño impecable	Visualizaciones claras, informativas, buen diseño	Visualizaciones básicas, información limitada, diseño simple	Visualizaciones pobres, poca información, diseño deficiente
Interpretación y contextualización	15 %	Interpretaciones profundas, contexto nacional integrado, insights valiosos	Interpretaciones apropiadas, contexto considerado, insights claros	Interpretaciones básicas, contexto limitado, pocos insights	Interpretaciones pobres, sin contexto, sin insights
Código R y reproducibilidad	10 %	Código elegante, completamente documentado, totalmente reproducible	Código funcional, bien documentado, reproducible	Código básico, documentación mínima, parcialmente reproducible	Código deficiente, poca documentación, no reproducible
Comunicación científica	5 %	Escritura clara y profesional, formato IEEE perfecto, estructura lógica	Escritura apropiada, formato IEEE correcto, estructura coherente	Escritura aceptable, formato IEEE básico, estructura simple	Escritura deficiente, formato incorrecto, estructura pobre
Total		**100 %**			

9. Especificaciones de entrega

Requisitos obligatorios para la entrega

Archivos a entregar:

1. **Archivo principal:** `ProyectoFinal_NombreEquipo.Rmd` (código fuente)
2. **Archivo compilado:** `ProyectoFinal_NombreEquipo.html` (documento final)
3. **Datos utilizados:** Todos los datasets en formato original
4. **Archivo CSS:** `estilo_ieee.css` (si se utiliza)
5. **Carpeta de imágenes:** Todas las imágenes externas utilizadas

Envío por correo electrónico:

- **Destinatario:** `jordyab00@gmail.com`
- **Remitente:** Correo institucional del coordinador del equipo
- **Asunto:** `ProyectoFinal_NombreEquipo_Tema`
- **Ejemplo:** `ProyectoFinal_Analiticos_AccidentesTransito`
- **Fecha límite:** Miércoles 5 de noviembre, 6:00 PM
- **Confirmación:** El profesor confirmará recepción dentro de 24 horas

En el cuerpo del correo incluir:

- Nombre del equipo
- Lista completa de integrantes (nombre completo y carné)
- Título del proyecto
- Breve descripción del tema (máximo 50 palabras)

10. Recursos de apoyo

Recursos disponibles

Librerías R recomendadas:

- `tidyverse` - Manipulación y análisis de datos
- `ggplot2` - Visualizaciones profesionales
- `plotly` - Gráficos interactivos
- `DT` - Tablas interactivas
- `knitr` - Generación de reportes
- `corrplot` - Matrices de correlación
- `sf` - Análisis espacial (si aplica)

Este proyecto de investigación representa la culminación del aprendizaje en análisis de datos, integrando competencias técnicas con pensamiento crítico para generar conocimiento relevante sobre la realidad nacional costarricense.