Análisis exploratorio de datos del Dataset "EVA" del IDEAM

Haessler Joan Ortiz Moncada

Universidad Distrital Francisco José de Caldas
Facultad de Ingeniería
Curso: BIG DATA

3 de septiembre de 2025

1. Descripción del Dataset

El conjunto de datos *EDA*, proporcionado por el IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia), contiene información sobre los cultivos agrícolas registrados en diferentes municipios de Colombia entre los años 2006 y 2018. En total, el dataset incluye aproximadamente 206 068 registros y diecisiete variables, las cuales se describen en la **??**:

2. Estadisticas básicas

Como forma de entender mejor el dataset, se realizó un análisis exploratorio de datos (EDA, por sus siglas en inglés) mediante la clase `Eda` desarrollada por mi e implementada en Python. Esta clase permite contar los registros, obtener los nombres y tipos de los campos, generar histogramas y diagramas de caja para variables numéricas, y crear tablas de contingencia para variables categóricas. Antes de presentar esta infromación, es importante mencionar que la clase incluye un método de limpieza de datos que elimina los registros con valores faltantes o inconsistentes, lo que redujo el número total de registros de 206 068 a 199 801. A continuación, se presentan algunas estadísticas básicas obtenidas del dataset:

- **Variable**: ha_semb
  - mean: 298.38
  - std: 1169.24
  - min: 0.00
  - 25 %: 10.00
  - 50 %: 38.00
  - 75 %: 160.00
  - max: 47 403.00

Nombre Campo	Significado	Tipo
cod_dpto	Código departamento	int64
departamento	Nombre departamento	str
cod_mun	Código municipio	int64
municipio	Nombre municipio	str
gr_cult	Grupo cultivo	str
Sub_gr_cult	Subgrupo cultivo	str
cultivo	Cultivo	str
des_reg_sist_prod	Desagregación regional y/o sistema productivo	str
year	Año	int64
prd	Periodo	str
ha_semb	Área sembrada (ha)	int64
ha_csda	Área cosechada (ha)	int64
t_prod	Produccion (t)	int64
t_ha_rend	Rendimiento (t/ha)	float64
st_fis_prd	Estado físico de producción	str
name_st	Nombre científico	str
cl_clt	Ciclo de cultivo	str

Cuadro 1: Campos del dataset EDA del IDEAM

- **Variable**: ha_csda
  - mean: 256.93
  - std: 994.66
  - min: 0.00
  - 25 %: 9.00
  - 50 %: 30.00
  - 75 %: 137.00
  - max: 38 600.00
- **Variable**: t_prod
  - mean: 2 874.78
  - std: 45 814.01
  - min: 0.00
  - 25 %: 36.00
  - 50 %: 150.00
  - 75 %: 4 546.00
  - max: 123 000.00
- **Variable**: t_ha_rend
  - mean: 9.27
  - std: 14.96

- min: 0.03
- 25 %: 1.50
- 50 %: 5.00
- 75 %: 11.50
- max: 246.00

Lógicamente, estas estadísticas solo son representativas para las variables numéricas del dataset. Estas estadísiticas permiten ver que existe una alta variabilidad en los datos, sin embargo, esto se explica por la presencia de outliers positivos altos, más no porque la varibilidad general de los datos sea alta.

Para las variables categóricas, se optó por mostrar una tabla de contingencia entre las variables **departamento** y **Sub_gr_cult** (y solo para AGUACATE, BANANO y CAFÉ), la cual se presenta en la **??**. Recordar que el presente informe no busca hacer un análisis exhaustivo, sino más bien ilustrar cómo realizar un Análisis Exploratorio de Datos (EDA por sus siglas en inglés) básico.

(a) Distribución por grupo etario

(b) Nivel educativo más alto alcanzado

Figura 1: Caracterización de la población encuestada

(a) Nivel de familiaridad con el término IA

(b) Capacidad de dar ejemplos de IA

Figura 2: Percepción de familiaridad y ejemplos de IA

(a) Confianza en los sistemas de IA

(b) Percepción sobre el uso de datos por parte del gobierno y las empresas

Figura 3: Confianza y percepción sobre uso de datos

Conclusiones

Tabla de contingencia (filtrada)			
Departamento	**AGUACATE**	**BANANO**	**CAFE**
AMAZONAS	8	2	0
ANTIOQUIA	674	286	1096
ARAUCA	28	0	1
ATLANTICO	2	0	0
BOLIVAR	89	0	20
BOYACA	159	62	471
CALDAS	271	62	300
CAQUETA	4	0	81
CASANARE	64	11	85
CAUCA	125	52	380
CESAR	182	13	228
CHOCO	29	165	12
CORDOBA	10	0	0
CUNDINAMARCA	268	179	817
GUAINIA	2	0	0
HUILA	332	228	420
LA GUAJIRA	127	24	122
MAGDALENA	0	72	48
META	110	22	155
NARIÑO	177	239	466
NORTE DE SANTANDER	171	63	422
PUTUMAYO	14	52	19
QUINDIO	141	171	144
RISARALDA	141	67	168
SAN ANDRES Y PROVIDENCIA	2	6	0
SANTANDER	281	106	852
SUCRE	39	0	0
TOLIMA	287	167	451
VALLE DEL CAUCA	417	475	468
VICHADA	5	3	0

Cuadro 2: Cantidad de cultivos por departamentos