Informe de Zonas Aptas para el Cultivo de Papa en Colombia

Fecha: 12/08/2025
Participantes: Jeronimo Cardona, Joshua Olaya, María Alejandra Quiroz

1. Introducción

Este informe presenta un análisis completo de las zonas en Colombia aptas y no aptas para el cultivo de papa, con base en el modelo de árbol de decisión previamente entrenado. El estudio considera variables como altitud y pH del suelo, que son factores determinantes en el rendimiento del cultivo.

El objetivo principal del proyecto es **apoyar a campesinos, productores agrícolas, entidades gubernamentales y organizaciones del sector agropecuario** en la toma de decisiones informadas sobre dónde establecer cultivos de papa con mayor probabilidad de éxito.

Contar con esta información puede contribuir a:

- Optimizar el uso de los recursos agrícolas.
- Reducir pérdidas económicas causadas por condiciones desfavorables.
- Promover prácticas agrícolas sostenibles.
- Facilitar la planificación y asignación de tierras aptas para el cultivo.
- Impulsar el desarrollo económico en regiones rurales.

Este análisis busca no solo brindar un respaldo técnico, sino también aportar al **fortalecimiento de la seguridad alimentaria** y al crecimiento del sector agrícola colombiano mediante el uso de herramientas de ciencia de datos y aprendizaje automático.

---

2. Metodología

2.1 Datos utilizados

- **Fuente de datos:** Dataset de clima para distintas regiones de Colombia.
- **Variables clave:**
  - Altitud (m.s.n.m.)
  - pH del suelo
- **Variable objetivo:** Aptitud para cultivo de papa (Apto / No apto)
- **Dataset del IDEAM:** /content/Cat_logo_Nacional_de_Estaciones_del_IDEAM_20250724.csv

2.2 Modelo aplicado

Se utilizó un **árbol de decisión** entrenado con el objetivo de clasificar zonas como aptas o no aptas. Las variables que arroja el modelo son las siguientes (contemplándose estás como las más importantes dentro de una gran cantidad de variables):

- Altitud entre 2.000 y 3.500 m.s.n.m. aumenta la probabilidad de ser apto.
- El pH del suelo entre 5.5 y 6.5 es ideal para el cultivo.
- El pH fue extraído de la Longitud y la Latitud extraídos del Dataset.

2.3 Evaluación del modelo

- **Precisión:** 94%
- **Matriz de confusión:**
- **Verdaderos Positivos (VP):** 222 → Zonas aptas correctamente identificadas como aptas.
- **Falsos Negativos (FN):** 28 → Zonas aptas clasificadas como no aptas.
- **Falsos Positivos (FP):** 28 → Zonas no aptas clasificadas como aptas.
- **Verdaderos Negativos (VN):** 264 → Zonas no aptas correctamente identificadas como no aptas.

---

3. Resultados

3.1 Lugares aptos

Según el modelo, las siguientes regiones presentan condiciones óptimas:

- Nariño (Pasto, Ipiales, Tuquerres)
- Boyacá (Duitama, Sogamoso, Tunja)
- Cundinamarca (Zipaquirá, Chocontá, Villapinzón)
- Antioquia (Santa Rosa de Osos, Entrerríos)

históricamente este tubérculo siempre ha estado presente en estas regiones desde épocas prehispánicas y organizadas "agrícolamente" para su producción en la época colonial, republicana y la actualidad.

3.2 Lugares no aptos

Regiones donde el pH o la altitud no cumplen con los requisitos:

- La Guajira (clima árido, baja altitud)
- Amazonas (suelo ácido y exceso de humedad)
- Costa Atlántica (altitud baja y altas temperaturas)

---

4. Visualización de datos

Durante el análisis exploratorio y evaluación del modelo se generaron las siguientes visualizaciones:

1. **Histograma del pH del suelo:** Muestra la distribución de los valores de pH con líneas de referencia para el rango óptimo (5.5 a 7.0).
2. **Histograma simple del pH del suelo:** Representación de la frecuencia de los valores de pH usando color naranja.
3. **Boxplots:**
    - Altitud (m.s.n.m.)
    - pH del suelo
    - Latitud
      Estos gráficos muestran la dispersión, valores atípicos y mediana de cada variable.
4. **Mapa de aptitud para cultivo de papa (scatterplot):** Puntos verdes = zonas aptas, puntos rojos = zonas no aptas, ubicados según coordenadas (longitud y latitud).
5. **Mapa interactivo en Folium:** Visualización geográfica con marcadores en las estaciones analizadas.
6. **Diagrama del árbol de decisión:** Representa gráficamente las reglas usadas por el modelo para clasificar.
7. **Gráfico de torta:** Proporción de zonas aptas vs no aptas.
8. **Gráfico de líneas:** Muestra la variación de aptitud a lo largo de valores de altitud y pH.
9. **Heatmap de correlación:** Muestra las relaciones entre las variables numéricas, destacando la correlación entre altitud y aptitud.
10. **Gráfico de barras:** Comparación entre cantidad de zonas aptas y no aptas.

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) es una fase fundamental en cualquier proyecto de ciencia de datos, ya que permite comprender la estructura, calidad y características de la información disponible antes de aplicar modelos o realizar predicciones. En este proyecto, el EDA Se desarrolló de la siguiente manera:

1. Carga de datos

Se verificó que la carga fuera correcta, revisando el número de registros y columnas, así como los primeros valores:

2. Evaluación de la calidad de los datos

Se realizó un diagnóstico inicial para detectar posibles problemas como:

- Valores nulos o ausentes.
- Datos duplicados.
- Formatos inconsistentes (fechas, texto, numéricos).
- Valores atípicos o fuera de rango.

3. Tratamiento de datos ausentes

Los valores ausentes se trataron según la naturaleza de cada variable:

1. En variables numéricas, se aplicó imputación con la media o mediana.

2. En variables categóricas, se imputó con la moda o una categoría especial como "Desconocido".

3. En casos donde el porcentaje de datos faltantes era muy alto, se consideró eliminar la columna.

4. Normalización de datos

Para garantizar que todas las variables numéricas tuvieran una escala comparable y evitar que algunas dominen sobre otras en los modelos, se aplicó normalización o estandarización según el caso:

- **Normalización Min-Max:** Escala los valores a un rango entre 0 y 1.
- **Estandarización Z-score:** Transforma los datos para que tengan media 0 y desviación estándar 1.

5. Resultados del EDA

Como resultado del EDA, el conjunto de datos, el conjunto de datos quedó limpio, consistente y listo para el análisis posterior y la aplicación de modelos de machine learning. Este proceso permitió:

- Reducir el riesgo de errores en el modelado.
- Asegurar que los datos fueran representativos.
- Mejorar la interpretabilidad de los resultados.

Evaluación del modelo

Para la clasificación de zonas aptas y no aptas para el cultivo de papa se utilizó un modelo de **árbol de decisión**.
La matriz de confusión generada fue la siguiente:

	Predicción: Apto	**Predicción: No apto**
Real: Apto	222	28
Real: No apto	28	264

Interpretación:

- **Verdaderos Positivos (VP):** 222 → Zonas aptas correctamente identificadas como aptas.
- **Falsos Negativos (FN):** 28 → Zonas aptas clasificadas como no aptas.
- **Falsos Positivos (FP):** 28 → Zonas no aptas clasificadas como aptas.
- **Verdaderos Negativos (VN):** 264 → Zonas no aptas correctamente identificadas como no aptas.

Métricas de evaluación:

- **Precisión (accuracy):** 0.94
- **Precision clase Apta:** 0.89
- **Recall clase Apta:** 0.89
- **F1-score clase Apta:** 0.89
- **Precision clase No apta:** 0.95
- **Recall clase No apta:** 0.95
- **F1-score clase No apta:** 0.95

Visualización de datos

Durante el análisis exploratorio y evaluación del modelo se generaron las siguientes visualizaciones:

11. **Histograma del pH del suelo:**
    Muestra la distribución de los valores de pH con líneas de referencia para el rango óptimo (5.5 a 7.0).
12. **Histograma simple del pH del suelo:**
    Representación de la frecuencia de los valores de pH usando color naranja.
13. **Boxplots:**
    - Altitud (m.s.n.m.)
    - pH del suelo
    - Latitud
      Estos gráficos muestran la dispersión, valores atípicos y mediana de cada variable.
14. **Mapa de aptitud para cultivo de papa (scatterplot):**
    Puntos verdes = zonas aptas, puntos rojos = zonas no aptas, ubicados según coordenadas (longitud y latitud).
15. **Mapa interactivo en Folium:**
    Visualización geográfica con marcadores en las estaciones analizadas.
16. **Diagrama del árbol de decisión:**
    Representa gráficamente las reglas usadas por el modelo para clasificar.
17. **Gráfico de torta:**
    Proporción de zonas aptas vs no aptas.
18. **Gráfico de líneas:**
    Muestra la variación de aptitud a lo largo de valores de altitud y pH.
19. **Heatmap de correlación:**
    Muestra las relaciones entre las variables numéricas, destacando la correlación entre altitud y aptitud.
20. **Gráfico de barras:**
    Comparación entre cantidad de zonas aptas y no aptas.

---

5. Conclusiones

1. Las zonas más aptas para el cultivo de papa en Colombia se concentran en la zona andina, especialmente entre 2.500 y 3.200 m.s.n.m.
2. El pH del suelo es un factor crítico: valores fuera del rango 5.5 - 6.5 reducen drásticamente la probabilidad de éxito.
3. El modelo logró identificar con alta precisión las zonas con condiciones ideales para el cultivo de papa, pero podría mejorarse incorporando variables como humedad, textura del suelo y temperatura promedio.
4. La herramienta desarrollada puede servir de apoyo a agricultores, instituciones y entidades gubernamentales para planificar siembras más eficientes.

---

6. Recomendaciones

- Ampliar la recolección de datos a más regiones para aumentar la robustez del modelo.
- Realizar validaciones en campo para confirmar la aptitud predicha.
- Actualizar periódicamente la base de datos para adaptarse a cambios ambientales y climáticos.
- Desarrollar una interfaz web o aplicación móvil que permita a los agricultores consultar la aptitud de su zona en tiempo real.

Anexo: Código del modelo y mapas generados disponibles en el repositorio de trabajo.