INFORME TÉCNICO

Ciclo de Vida de un Proyecto de Machine Learning para el Análisis de Área Cosechada

Descripción:

Este informe documenta el ciclo de vida de un proyecto de Machine Learning centrado en analizar datos de área cosechada, desde la detección del problema hasta la obtención de resultados y métricas de evaluación. El análisis se enfoca en identificar patrones y tendencias en la superficie cultivada para apoyar la toma de decisiones en el sector agrícola.

Campistas Talento TECH.
Equipo N.º 1

Informe Técnico: Ciclo de Vida de un Proyecto de Machine Learning para el Análisis de Área Cosechada

1. Introducción

El ciclo de vida de un proyecto de Machine Learning (ML) abarca una serie de fases que permiten abordar un problema real mediante el uso de algoritmos predictivos y análisis de datos. Este informe detalla las etapas clave seguidas para desarrollar un modelo de ML enfocado en analizar datos de área cosechada. Este análisis permite identificar patrones en la producción agrícola, apoyando la toma de decisiones estratégicas en el sector.

El objetivo principal de este proyecto es entender las tendencias y variaciones en el área cultivada y explorar cómo estos datos pueden usarse para predicciones futuras.

2. Ciclo de Vida de Proyectos de Machine Learning

El ciclo de vida del proyecto incluye las siguientes fases:

1. Detección del Problema
2. Identificación de Datos y Stakeholders
3. Análisis Exploratorio de Datos (EDA)
4. Preprocesamiento y Limpieza de Datos
5. Entrenamiento y Selección del Modelo
6. Validación y Ajuste
7. Despliegue y Monitoreo.

3. Selección del Contexto y Problema Específico

Contexto

En este proyecto, el contexto es la agricultura. La variabilidad en el área cosechada afecta directamente la producción agrícola, la planificación de recursos y la rentabilidad.

Problema Específico

Detectar tendencias y patrones en el área cosechada, permitiendo entender los factores que influyen en la variación del área y apoyar a los agricultores y tomadores de decisiones a optimizar recursos y planificar estrategias futuras.

4. Identificación de Datos y Stakeholders

Identificación de Datos

Los datos fueron recopilados de registros agrícolas históricos y consisten en información sobre la superficie cosechada en diversas regiones, así como el rendimiento de las

cosechas en diferentes períodos de tiempo. No se incluyen factores climáticos ni el uso de insumos, entre otros aspectos.

Se incluyen variables como:

- Superficie total cosechada en hectáreas
- Rendimiento agrícola (producción por hectárea)

Estos datos son fundamentales para comprender el comportamiento de las cosechas y la influencia de diferentes factores sobre el área cosechada.

Identificación de Stakeholders

Los stakeholders clave en este proyecto incluyen:

- **Agricultores**: Usuarios finales de las predicciones, quienes pueden usar los resultados para planificar sus actividades agrícolas.
- **Tomadores de Decisiones y Políticas Agrícolas**: Interesados en los patrones y predicciones para crear políticas que apoyen la sostenibilidad agrícola y optimicen el uso de recursos.
- **Instituciones de Investigación Agrícola**: Pueden utilizar los resultados para mejorar sus estudios en el área de ciencia de datos y agricultura.
- **Financieros e Inversionistas**: Aquellos interesados en el rendimiento del sector agrícola, quienes utilizan los datos para planificar inversiones.

Cada grupo de stakeholders está interesado en cómo los resultados pueden ayudar a optimizar los rendimientos agrícolas, la gestión de insumos y la planificación de recursos.

5. Análisis Exploratorio de Datos (EDA)

5.1 Carga de Datos

Para este análisis, los datos fueron cargados desde un archivo de datos abiertos proporcionado por el Ministerio de Cultura y descargado por los campistas desde la web. Este archivo contiene información detallada sobre el área cosechada y rendimientos en varias regiones a lo largo de diferentes años.

5.2 Evaluación de Calidad de Datos

- **Valores Faltantes**: Se identificaron datos faltantes en algunas de las columnas críticas. Estos valores fueron tratados mediante la imputación con el promedio de cada columna.
- **Valores Atípicos**: Algunos valores presentaban variaciones extremas que fueron revisadas y, cuando fue necesario, ajustadas.

5.3 Tratamiento de Datos Ausentes

Para manejar los datos faltantes, se implementó una estrategia de imputación, utilizando la media y mediana para garantizar la consistencia en el análisis.

5.4 Normalización de Datos

Los datos fueron normalizados para asegurar que todas las variables estuvieran en un rango comparable, especialmente aquellas con escalas muy diferentes.

5.5 Análisis Univariado

- Se realizaron histogramas para observar la distribución de cada variable. Esto ayudó a identificar distribuciones sesgadas y a visualizar tendencias en el área cultivada.

5.6 Análisis Bivariado

- Se realizó un análisis de correlación para determinar la relación entre el área cosechada y otras variables como el rendimiento y los insumos utilizados. Los diagramas de dispersión se usaron para visualizar estas relaciones.

5.7 Análisis Multivariado

- Utilizando gráficos de pares y análisis de componentes principales (PCA), se exploraron relaciones más complejas entre múltiples variables. Esto ayudó a identificar patrones ocultos y a reducir la dimensionalidad de los datos para análisis posteriores.

5. Desarrollo del Modelo y Métricas de Evaluación

Para evaluar el desempeño del modelo, se utilizaron las siguientes métricas:

Métrica	Resultado
R^2 en conjunto de prueba	0,82
R^2 en Validación Cruzada (fold 1)	0,9326
R^2 en Validación Cruzada (fold 2)	0,9453
R^2 en Validación Cruzada (fold 3)	0,9398
R^2 en Validación Cruzada (fold 4)	0,7799
R^2 en Validación Cruzada (fold 5)	0,8011
Media de R^2 en Validación Cruzada	0,8798

- **Resultados de Métricas Individuales**:
    - Valores obtenidos de la predicción: [30.99, 325.99, 4.87, ..., 183.57, 3.30, 329.90]
- **R^2 (Coeficiente de Determinación)**: Mide qué tan bien el modelo predice el valor observado. En este caso, el valor obtenido fue de **0.82**, indicando una buena capacidad de ajuste del modelo al conjunto de prueba.
- **Validación Cruzada**:
    - Los puntajes de validación cruzada de R^2 fueron: [0.9326, 0.9453, 0.9398, 0.7799, 0.8011].
- **Media de R^2 en Validación Cruzada**: **0.88**, lo que indica que el modelo tiene un rendimiento estable y generaliza bien en los datos de validación.

Estos resultados indican un buen desempeño del modelo, lo cual respalda su uso para predecir la superficie de área cosechada y sugiere su aplicabilidad en la planificación de recursos en el sector agrícola.

7. Conclusiones

En este análisis preliminar de datos de área cosechada, se identificaron patrones estacionales y factores que influyen en la variabilidad de la superficie cultivada. Los resultados obtenidos serán de utilidad para el desarrollo de modelos predictivos que permitan prever cambios en el área cosechada, facilitando una planificación más efectiva en el sector agrícola.

Este informe presenta la primera fase de un proyecto de Machine Learning orientado al análisis y predicción de tendencias agrícolas. En futuras fases, se espera desarrollar y evaluar diferentes modelos para mejorar la precisión de las predicciones.

8. Participantes

Investigador Principal:

Científico de Datos:

Ingeniera de Machine Learning: Marian Betancourt Uribe

Analista de Datos:

Coordinador de Proyecto:

9. Bibliografía

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.