Nadia Arellano - Ana Karina Muñoz
Ingenieras en Informática

Predicción de Lluvias mediante Machine Learning con metodología CRISP-DM

Descripción del Proyecto

Este proyecto aplica técnicas de minería de datos sobre el dataset meteorológico Weather Australia para predecir lluvias y estimar temperaturas máximas con el fin de mejorar la planificación agrícola en cultivos de trigo. Se abordó el problema utilizando la metodología CRISP-DM, siguiendo todas sus fases para preparar, modelar y evaluar los datos mediante herramientas de Python en Jupyter Notebook.

Arquitectura Analítica y Técnicas Utilizadas

- Preprocesamiento y limpieza de datos (imputación por mediana/moda, detección de outliers, estandarización con StandardScaler).
- Transformación de variables categóricas con LabelEncoder y creación de nuevas variables como Season.
- Modelos de regresión: Regresión Lineal Múltiple y Árbol de Regresión para estimar MaxTemp.
- Modelos de clasificación: Árbol de Decisión y Naive Bayes para predecir RainTomorrow.

Casos de Uso Analíticos

- ¿Lloverá mañana en una zona específica?
- ¿Qué temperatura máxima se puede esperar en determinada estación o época del año?
- ¿Cómo se comportan los patrones de lluvia por estado o región?

Nadia Arellano - Ana Karina Muñoz
Ingenieras en Informática

Resultados de Evaluación de Modelos

La evaluación de los modelos desarrollados se realizó considerando tanto técnicas de regresión como de clasificación, siguiendo métricas específicas de rendimiento.

Modelos de Regresión

Para la predicción de la temperatura máxima (MaxTemp), se compararon dos enfoques: Regresión Lineal Múltiple y Árbol de Regresión**.** Como se observa en la **Figura 1,** el modelo de Regresión Lineal Múltiple obtuvo mejores resultados, presentando un menor MAE (0.329)**,** menor MSE (0.177) y un coeficiente de determinación R^2 más alto (0.821) en comparación al Árbol de Regresión.

Métricas	Modelos	
	Regresión Lineal Múltiple	Árbol de Regresión
MAE (Error Absoluto Medio)	0.329	0.370
MSE (Error Cuadrático Medio)	0.177	0.223
R^2 (Coeficiente de Determinación)	0.821	0.774

Figura 1. Comparación de métricas de evaluación para los modelos de regresión. Estos resultados respaldan la selección de la Regresión Lineal Múltiple como modelo adecuado para estimar la temperatura máxima en escenarios agrícolas.

Modelos de Clasificación

En la predicción de la probabilidad de lluvia al día siguiente (RainTomorrow), se evaluaron los modelos de Árbol de Decisión y Naive Bayes. Según se presenta en la **Figura 2**, el Árbol de Decisión logra un mejor equilibrio entre métricas como precisión global (83%), AUC (82%), y F1-Score (81%), superando consistentemente al modelo de Naive Bayes.

Métricas	Modelos	
	Árbol de Decisión	Naive Bayes
AUC (Área bajo la curva ROC)	82%	81%
Precisión Global (Accuracy)	83%	81%
Precisión Clase 0	85%	85%
Precisión Clase 1	70%	44%
F1-Score	81%	80%

Figura 2. Comparativa de desempeño entre modelos de clasificación para la predicción de RainTomorrow. El Árbol de Decisión logra un mejor equilibrio, por lo que se selecciona como el modelo final en este proyecto.

Dado el rendimiento general y la importancia de mantener un balance entre las clases (días de lluvia y días sin lluvia), el Árbol de Decisión fue seleccionado como el modelo final general.

Visualización y Resultados

- Generación de archivo CSV con predicciones, reversión de escalado y decodificación de variables.
- Creación de la columna 'Predicción Correcta' para evaluación individual.
- Insights que ayudan a agricultores a optimizar siembra, reducir pérdidas y planificar en función del clima.

Herramientas Tecnológicas

- Jupyter Notebook
- Python 3.8
- Bibliotecas: pandas, numpy, matplotlib, seaborn, scikit-learn
- Looker Studio

Impacto Profesional

Este proyecto demuestra competencias en preparación avanzada de datos, modelado predictivo, evaluación rigurosa de modelos y generación de insights accionables en sectores productivos como la agricultura. La integración de predicciones climáticas con decisiones estratégicas permite mejorar el rendimiento de los cultivos y la eficiencia del uso de recursos.