Escuela de Informática
Y Telecomunicaciones

DuocUC

Examen Transversal

“Dataset Clima Australia”

04-07-2025

Integrantes:

Nicolás Oses
Agustín Quezada
Danilo Morales

Asignatura:

Minería de Datos

Docente:

Marco Japke

Tabla de contenido

Introducción

Objetivo del informe:

Este informe presenta el desarrollo de un proyecto de minería de datos cuyo objetivo principal es analizar y extraer características del clima en Australia, con el fin de predecir la temperatura máxima diaria y la probabilidad de lluvia al día siguiente en distintas ubicaciones del país. Para ello, se utiliza un conjunto de datos meteorológicos históricos. A lo largo de distintas etapas que incluyen análisis exploratorio, procesamiento de datos, modelado y evaluación, se busca generar información valiosa que contribuya a una mejor toma de decisiones en áreas sensibles al clima, como la agricultura, la logística y la planificación urbana.

Contexto del problema:

En Australia, el comportamiento del clima es altamente variable y, en muchas zonas, las lluvias son un factor crítico que afecta directamente las actividades productivas. La incertidumbre respecto a si lloverá o no puede traducirse en costos económicos, desperdicio de recursos o incluso afectación a la seguridad de las personas. Frente a este escenario, el uso de técnicas de minería de datos permite transformar los datos históricos del clima en información predictiva confiable, apoyando así decisiones informadas por parte de organizaciones públicas y privadas.

Este trabajo se enmarca en la asignatura de Minería de Datos y se desarrolló siguiendo una metodología estructurada que incluyó desde la exploración y preparación de los datos hasta el entrenamiento de modelos predictivos y el descubrimiento de patrones relevantes.

Metodología Empleada

Enfoque de trabajo:

Para el desarrollo de este proyecto se aplicó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente utilizada en el área de la ciencia de datos por su estructura clara y adaptable. Esta metodología divide el trabajo en seis etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En este caso, se abordaron las cinco primeras etapas, ya que no se contempló la implementación directa del modelo en un entorno productivo.

Esta estructura permitió abordar el problema de forma ordenada, desde el análisis inicial de la situación hasta la evaluación de los resultados obtenidos a partir de los modelos predictivos. El uso de CRISP-DM facilitó la toma de decisiones en cada fase del proyecto, asegurando la trazabilidad y coherencia del proceso.

Análisis previo del problema y fuentes de datos:

El problema principal planteado fue la necesidad de predecir si lloverá al día siguiente (RainTomorrow), además de predecir la temperatura máxima del día (MexTemp) en base a variables climáticas como temperatura, humedad, presión atmosférica, localidad y estación del año. Para ello, se utilizó un dataset público de clima de Australia, el cual contiene datos meteorológicos diarios registrados en distintas estaciones a lo largo del país durante varios años (2007-2017).

Este conjunto de datos incluyó variables numéricas (como temperatura máxima y mínima, humedad, viento, presión) y categóricas (como dirección del viento, localidades, nubosidad etc.), permitiendo un análisis completo y variado del comportamiento climático.

Selección de técnicas y modelo conceptual:

Considerando que la variable a predecir de clasificación (RainTomorrow) es de tipo binaria (“Yes” o “No”), se optó por aplicar técnicas de clasificación supervisada. En específico, se trabajó con modelos como:

- Regresión Logística con Cross Validation
- Árboles de Decisión (Decision Tree)
- KNN de clasificación

Mientras que para la variable a predecir de regresión (MaxTemp), se trabajó con modelos como:

- Regresión Lineal Ridge
- Árboles de Decisión de regresión
- KNN de regresión

Estas técnicas se seleccionaron por su facilidad de interpretación, eficiencia en la predicción de variables categóricas y numéricas, además de ser ampliamente utilizadas en problemas similares. El modelo conceptual del proyecto consistió en utilizar estas técnicas para generar predicciones basadas en el análisis conjunto de las distintas variables del clima disponibles en el dataset

Etapa 1: Preparación de Datos

Limpieza y preprocesamiento:

Al explorar el dataset por primera vez, se detectó la existencia de múltiples valores nulos (NaN), tanto en variables numéricas como categóricas. Para asegurar la calidad del análisis, se aplicaron las siguientes acciones de limpieza:

Eliminación de columnas con alta cantidad de datos faltantes, como Evaporation y Sunshine, ya que más del 40% de sus registros estaban vacíos.

Imputación de valores faltantes en variables relevantes como MinTemp, MaxTemp y Humidity mediante el uso de KNN Imputer.

Conversión de la variable objetivo RainTomorrow a valores binarios, asignando 1 a “Yes” y 0 a “No”, lo que facilitó su análisis con modelos de clasificación.

Transformación de variables categóricas (por ejemplo, Localidad y Estacion del año) a variables numéricas mediante codificación (one-hot encoding o label encoding, según correspondía).

Tratamiento de outliers por localidad y estación del año nos permitió abordar y mejorar observaciones que se alejan del patrón general del conjunto de datos, para ello remplazamos los valores outliers por la mediana.

Escalamiento de los datos para ajustar todas las variables numéricas a una escala común entre 0 y 1.

Análisis exploratorio de los datos

Antes de aplicar modelos, se realizó un análisis exploratorio para comprender mejor la estructura y comportamiento del dataset:

Se visualizaron distribuciones de temperatura, humedad y lluvia, observando tendencias estacionales.

Se identificaron correlaciones entre variables, destacando que la humedad al atardecer (Humidity3pm) tenía una fuerte relación con la probabilidad de lluvia al día siguiente.

Se usaron gráficos como histogramas, mapas de calor y diagramas de caja (boxplots) para detectar valores atípicos (outliers) y evaluar la dispersión de los datos.

Este análisis preliminar permitió seleccionar las variables más relevantes y definir una base sólida para aplicar los modelos predictivos.

Etapa 2: Aplicación de Técnicas de Minería de Datos

Técnicas utilizadas:

Dado que el objetivo del proyecto era predecir una variable binaria (lloverá o no) y la temperatura máxima del día, se optó por aplicar técnicas de clasificación y regresión supervisada. Estas técnicas permiten construir modelos que, a partir de un conjunto de características conocidas, puedan predecir una categoría de salida.

Las técnicas utilizadas para clasificación fueron:

Regresión Logística: método estadístico que estima la probabilidad de que ocurra un evento, útil cuando la variable objetivo es dicotómica.

Árbol de Decisión: modelo que toma decisiones dividiendo los datos en función de reglas simples, lo que lo hace fácil de interpretar.

KNN de clasificación: algoritmo supervisado que asigna etiquetas a nuevas observaciones en función de la similitud con sus vecinos más cercanos.

Y para regresión fueron:

- Regresión Lineal Ridge
- Árboles de Decisión de regresión
- KNN de regresión
- Implementación

Para aplicar los modelos, se dividió el conjunto de datos en dos partes:

- **Conjunto de entrenamiento** (70% de los datos): utilizado para construir los modelos.

- **Conjunto de prueba** (30%): utilizado para validar el rendimiento de cada modelo sobre datos no vistos.

Además, para los modelos de clasificación de dividió el conjunto de datos de manera estratificada.

Cada modelo fue ajustado utilizando parámetros por defecto en una primera etapa, y luego se realizaron pequeñas mejoras mediante validación cruzada y GridSearchCV.

Además, se compararon los resultados obtenidos por cada técnica para determinar cuál ofrecía la mejor capacidad predictiva en este caso.

Etapa 3: Modelos Predictivos y Evaluación

Desarrollo de los modelos:

Una vez que los datos fueron preparados y divididos en conjuntos de entrenamiento y prueba, se procedió al entrenamiento de distintos modelos de clasificación y regresión supervisada.

Evaluación del rendimiento clasificación

Para evaluar la efectividad de cada modelo, se utilizaron las siguientes métricas:

- **Accuracy (precisión):** porcentaje de aciertos del modelo sobre el total de casos.
- Matriz de confusión: tabla que permite observar cuántos casos fueron correcta o incorrectamente clasificados, tanto positivos como negativos.
- **Recall y F1-score:** métricas complementarias que ayudan a evaluar el rendimiento en clases desbalanceadas.
- **ROC AUC:** Evalúa la calidad del modelo en la clasificación binaria, analizando su capacidad para distinguir entre clases positivas y negativas.

Los resultados fueron los siguientes:

Modelo	Accuracy	Observación
Regresión Logística	~75%	Buen desempeño base
Árbol de Decisión	~86%	Mejor rendimiento general
KNN clasificación	~83%	Cercano al rendimiento de árbol de decisión.

Evaluación del rendimiento regresión

Para evaluar la efectividad de cada modelo, se utilizaron las siguientes métricas:

- **MAE (Mean Absolute Error):** mide el promedio de los errores absolutos entre las predicciones y los valores reales. Indica, en promedio, cuánto se desvía el modelo del valor real.
- **MSE (Mean Squared Error**): calcula el promedio de los errores al cuadrado. Penaliza más fuertemente los errores grandes, lo que permite identificar desviaciones importantes en las predicciones.
- **RMSE (Root Mean Squared Error):** es la raíz cuadrada del MSE y mantiene las unidades originales de la variable objetivo, lo que facilita la interpretación del error promedio de las predicciones.
- **R^2 (Coeficiente de Determinación):** indica qué proporción de la variabilidad de la variable dependiente puede explicarse por el modelo. Su valor va de 0 a 1, donde valores cercanos a 1 indican un mejor ajuste del modelo.

Los resultados fueron los siguientes:

Modelo	RMSE	Observación
Regresión Ridge	~0.0250	Mejor rendimiento
Árbol de Decisión regresión	~0.0302	Menor precisión que los modelos lineales
KNN regresión	~0.0272	Errores algo mayores que los modelos lineales, pero menores que el Árbol de decisión.

Etapa 4: Descubrimiento de Patrones con Clustering

Identificación de patrones relevantes

El Aprendizaje No Supervisado es una rama del aprendizaje automático que se enfoca en descubrir patrones ocultos en los datos sin la necesidad de etiquetas o respuestas predefinidas. Dentro de este enfoque, el Clustering es una técnica fundamental que agrupa datos en conjuntos según su similitud.

Para ello usamos las siguientes variables corresponden a mediciones meteorológicas claves tomadas a las 9 de la mañana y a las 3 de la tarde, lo que permite capturar tanto las condiciones atmosféricas matutinas como las de la tarde**.: 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Temp9am', 'Temp3pm'.**

Aplicación del modelo:

Se utilizó el modelo K-Means junto con reducción de dimensionalidad mediante PCA, ajustándolo en función de lograr la menor inercia posible y la mejor calidad de agrupación, evaluada mediante el coeficiente de Silhouette y el método del Codo (Elbow). A partir de estos criterios, se determinó que el valor óptimo de k es 5.

PCA con 3 componentes principales:

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de dimensionalidad utilizada en el aprendizaje automático. Su objetivo es transformar un conjunto de variables en un nuevo grupo de variables llamadas componentes principales, que son ortogonales entre sí y capturan la mayor cantidad de variabilidad posible en los datos. Al reducir la cantidad de dimensiones, PCA facilita el análisis y el entrenamiento de modelos, eliminando redundancias y destacando las características más relevantes.

Se obtuvo un ~90% de la varianza explicada con 3 componentes principales. A continuación se explican cada componente:

- **PC1**: Influenciado por los niveles de humedad tanto en la mañana como en la tarde, y en menor medida por la temperatura vespertina. Esto sugiere que este componente representa una especie de "eje de humedad ambiental", posiblemente asociado a la presencia de nubes, lluvia, o sensación térmica.
- **PC2:** Este componente combina medidas de presión atmosférica (mañana y tarde) con la humedad vespertina. Parece capturar condiciones de estabilidad atmosférica relacionadas con sistemas de alta o baja presión, pero también reflejando el grado de saturación del aire por la tarde.
- **PC3:** Este componente está dominado por la temperatura en la mañana, junto con valores de presión. Refleja condiciones térmicas y de presión en las primeras horas del día, lo que podría influir en cómo se desarrollará el clima durante el resto del día.

Análisis de las agrupaciones obtenidas con K-Means + PCA.

Cluster 0: Clima cálido y húmedo, representa días calurosos con humedad considerable, típicos de climas subtropicales.

- Humedad: 70% AM, 59% PM (moderadamente alta)
- Presión: Baja (≈1013 hPa)
- Temperatura: Alta (23°C AM, 26°C PM)

Cluster 1: Clima muy caluroso y seco, grupo más extremo en términos de calor y sequedad. Representa condiciones similares a las de olas de calor o zonas áridas en verano.

- **Humedad:** Muy baja (42% AM, 22% PM)
- **Presión:** Baja (≈1011 hPa)
- **Temperatura:** Muy alta (23°C AM, 31°C PM)

Cluster 2: Clima frío y muy húmedo, son días fríos y saturados de humedad, posiblemente asociados a lluvias persistentes, niebla o cielos muy cubiertos. Común en invierno u otoño húmedo.

- Humedad: Muy alta (87% AM, 72% PM)
- Presión: Media (≈1014 hPa)
- Temperatura: Muy baja (13°C AM, 15.7°C PM)

Cluster 3: Clima frío y estable que Corresponden a días fríos pero estables, dominados por sistemas de alta presión. Podrían ser días despejados, secos y con aire frío, típicos de invierno.

- **Humedad:** Alta (82% AM, 59% PM)
- **Presión:** Muy alta (≈1026–1024 hPa)
- **Temperatura:** Baja (11°C AM, 16°C PM)

Cluster 4: Clima templado y seco son días agradables y secos, probablemente comunes en primavera u otoño. Condiciones confortables con buen tiempo.

- **Humedad:** Moderada-baja (63% AM, 40% PM)
- **Presión:** Alta (≈1019–1017 hPa)
- **Temperatura:** Moderada (15°C AM, 21°C PM)

Documentación del proceso y conclusiones

Documentación del proceso completo

Durante el desarrollo de este proyecto se siguió un enfoque sistemático basado en la metodología CRISP-DM. Cada etapa fue documentada cuidadosamente, desde la obtención del dataset hasta la evaluación de los modelos finales. Se utilizaron notebooks de Jupyter para organizar el trabajo de manera clara y reproducible, registrando los pasos, decisiones y resultados intermedios en cada fase.

El proceso incluyó:

- ✓ Revisión y limpieza de los datos meteorológicos históricos.
- ✓ Análisis exploratorio y visualización de relaciones entre variables.
- ✓ Aplicación de modelos de clasificación, regresión y clustering.
- ✓ Evaluación comparativa y selección del modelo más efectivo.

Esta documentación no solo permitió dar seguimiento al avance del proyecto, sino también facilitar su comprensión y análisis por parte de terceros.

Desafíos encontrados

Uno de los principales desafíos fue la presencia de datos faltantes y desbalance en la variable objetivo, lo que exigió un trabajo cuidadoso en el preprocesamiento. También fue importante tomar decisiones respecto a qué columnas conservar, transformar o eliminar para mantener la calidad del análisis sin perder información valiosa.

Insights adicionales:

Además de las predicciones, el proyecto permitió identificar comportamientos repetitivos en ciertas condiciones climáticas que podrían extenderse a otros contextos o regiones. Por ejemplo, el patrón de humedad alta al atardecer como predictor de lluvia podría explorarse en zonas con climas similares.

Recomendaciones y próximos pasos

A partir de los resultados obtenidos, se recomienda considerar la implementación práctica del modelo de predicción, especialmente en sectores como la agricultura, donde anticipar lluvias puede optimizar recursos y evitar pérdidas.

También se sugiere continuar recolectando datos climáticos actualizados para mantener el modelo vigente y mejorar su precisión. Además, incorporar nuevas variables (como calidad del aire, eventos extremos o imágenes satelitales) podría enriquecer el análisis.

Como próximo paso, se propone desarrollar un flujo automatizado que integre el modelo con fuentes de datos en tiempo real, permitiendo generar alertas o predicciones diarias de forma autónoma.