Penquist RainForecast: Modelo predictivo de lluvias en Concepción, Chile

Bastian Arriagada Quero, Diego Vargas Gómez
Universidad del Bio-Bio, Collao #1202
bastian.arriagada2201@alumnos.ubiobio.cl, diego.vargas2001@alumnos.ubiobio.cl

Resumen
El proyecto Penquist RainForecast se centra en desarrollar un modelo predictivo para estimar la probabilidad de lluvia diaria en Concepción usando algoritmos de machine learning. Trabajando con datos históricos meteorológicos recolectados de la estación Carriel Sur de Talcahuano, este modelo integrará variables como cobertura nubosa, humedad relativa, temperatura, presión atmosférica y más. La precisión del modelo será evaluada mediante métricas estándar como lo es la precisión, además, se implementarán gráficos y matrices para facilitar la interpretación de los resultados. Este modelo tiene aplicaciones potenciales en agricultura, gestión de recursos hídricos y planificación urbana.

Palabras claves
Concepción, predicción de lluvia, meteorología, machine learning, análisis de datos.

Metadatos

Nr	**Descripción de metadatos**	
C1	Versión actual del código	v1.0
C2	Enlace permanente al repositorio del código	https://github.com/Diegariio/Modelos-de-ML-DL
C3	Enlace a una cápsula reproducible	No aplica.
C4	Licencia del código	No aplica.
C5	Sistema de control de versiones	Git
C6	Lenguajes de programación, herramientas y servicios usados	Jupyter notebook, python, excel.
C7	Requisitos de compilación, entornos operativos y dependencias	Python 3.12, Pandas, Scikit-Learn, Matplotlib, tensorflow, keras.
C8	Enlace a la documentación del desarrollador o manual	No existe
C9	Correo electrónico de soporte para preguntas	bastian.arriagada2201@alumnos.ubiobio.cl diego.vargas2001@alumnos.ubiobio.cl

1. Motivación e Importancia

La predicción precisa de la lluvia es crucial para diversas aplicaciones como la agricultura, la gestión de recursos hídricos y la planificación urbana. El proyecto Penquist RainForecast se propone desarrollar un modelo predictivo utilizando algoritmos de machine learning para estimar si lloverá o no en la ciudad de Concepción, Chile. La motivación principal para desarrollar este software es mejorar la precisión de las predicciones meteorológicas locales, lo que puede ayudar a los agricultores a planificar mejor sus cultivos, a los gestores de recursos hídricos a optimizar el uso del agua, y a las autoridades urbanas a prepararse mejor para eventos climáticos extremos.

El software aborda el problema de la predicción de la lluvia mediante la integración de múltiples variables meteorológicas como la cobertura nubosa, la humedad relativa, la temperatura, la presión atmosférica, entre otras más.

En el futuro, RainForecast contribuirá al descubrimiento científico al proporcionar una herramienta accesible y precisa para la predicción meteorológica, que puede ser utilizada por estudiantes tesistas e investigadores interesados en el área del clima. La mejora en la precisión de las predicciones puede llevar a una mejor comprensión de los patrones climáticos locales y sus impactos.

2. Descripción del modelo

Este proyecto tiene como objetivo recopilar y analizar datos climáticos para predecir si lloverá o no en determinadas fechas. Utiliza un dataset que abarca desde 1970 hasta 2023 con datos de los meses de abril, mayo, junio, julio, agosto y septiembre. Cada fila del dataset incluye variables meteorológicas diarias como la cantidad de cielo cubierto media, humedad relativa media, temperatura media, dirección y fuerza del viento predominante, presión QFE media, y un indicador de si llovió o no (Extraído de la página de la Dirección Meteorológica de Chile). El proyecto emplea técnicas de Machine Learning para entrenar modelos predictivos con estos datos.

2.1. Arquitectura de los modelos*:*

La arquitectura del software se puede dividir en varias etapas:

- **Recopilación de datos**: Se recolectan datos climáticos diarios desde la Dirección metereologica de Chile utilizando métodos de Web Scraping y se organiza todo en un Excel. Posteriormente se hace la conversión de un archivo. xlss a un .csv para la utilización del dataset.

- **Preprocesamiento de datos**: Se limpia y prepara el dataset, separando las variables predictoras del objetivo (si llovió o no).

- **Separación del dataset**: Se divide el dataset en conjunto de entrenamiento (80%) y conjunto de prueba (20%).

- **Entrenamiento de modelos**: Se entrenan tres tipos de modelos: Random Forest, Árbol de Decisión y Redes Neuronales.

- **Evaluación de modelos**: Se evalúan los modelos utilizando el conjunto de prueba para determinar su precisión y capacidad predictiva.

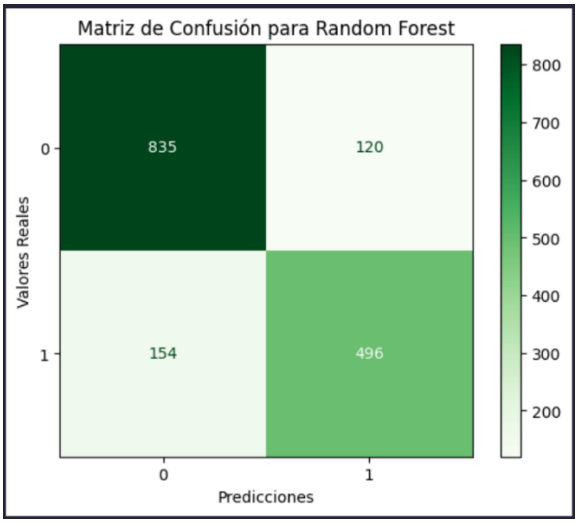
2.2. Funcionalidades del modelo*:*

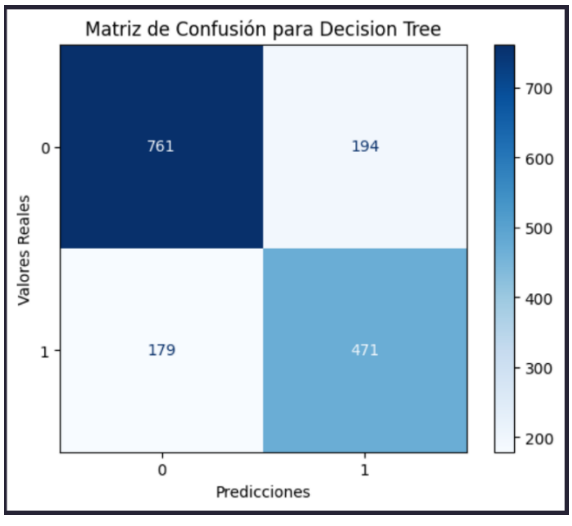
RainForecast ofrece las siguientes funcionalidades principales:

- **Evaluación de Modelos**: Calcula y muestras métricas de rendimiento (****) para evaluar la precisión de los modelos entrenados.

- **Visualización de Resultados**: Ofrece herramientas para visualizar las predicciones de lluvia y analizar el rendimiento de los modelos mediante gráficos y tablas.

3. Ejemplos ilustrativos (Matriz de confusión)





4. Impacto

Este proyecto, aunque es un modelo de entrenamiento y no un software de gran escala, tiene varias implicaciones importantes en el ámbito académico y potencialmente en el sector meteorológico.

4.1 Nuevas preguntas de investigación

El proyecto abre la puerta a varias nuevas preguntas de investigación, tales como:

- ¿Cómo han cambiado los patrones de lluvia a lo largo de las décadas y qué factores climáticos influyen más en estos cambios?
- ¿Qué técnicas de Machine Learning pueden mejorar la precisión de las predicciones meteorológicas a partir de datos históricos?
- ¿Cuáles son las variables más influyentes en la predicción de lluvia y cómo pueden ser mejor capturadas y utilizadas en modelos predictivos?

4.2 Cambios en la práctica diaria de los usuarios

El proyecto proporciona a los estudiantes una experiencia práctica en la aplicación de técnicas de Machine Learning a datos reales, mejorando sus habilidades y comprensión en el campo, además, fomenta la colaboración entre estudiantes al proporcionar una plataforma común para el análisis de datos climáticos y el desarrollo de modelos predictivos.

4.3 Uso del software dentro y fuera del grupo de usuarios previstos

Dado que este proyecto es parte de una asignatura universitaria de introducción al Machine Learning, su uso se limita principalmente a estudiantes universitarios, ya que el proyecto se utiliza como herramienta educativa para enseñar técnicas de Machine Learning aplicadas a datos climáticos.

4.4 Posibles contribuciones

- **Consultoría climática:** Los modelos desarrollados podrían ser utilizados por empresas de consultoría climática para mejorar sus predicciones y análisis.
- **Desarrollo de software meteorológico:** Podría inspirar el desarrollo de software más avanzado para la predicción del clima, utilizando técnicas similares de Machine Learning.

Aunque este proyecto es de naturaleza académica y no tiene un uso comercial directo, sus contribuciones al campo de la predicción climática y la educación en Machine Learning son significativas. Proporciona una base sólida para futuras investigaciones y aplicaciones más avanzadas.

5. Conclusiones y comentarios

En el análisis realizado a los modelos para predecir la lluvia, se entrenaron tres: Árbol de Decisión, Random Forest y Redes Neuronales. Los resultados obtenidos muestran que el modelo de Random Forest obtuvo la mayor precisión, seguido por las Redes Neuronales y finalmente el Árbol de Decisión. Estos resultados indican que los modelos más complejos y avanzados, como el Random Forest, son más efectivos para capturar las relaciones complejas y no lineales presentes en los datos meteorológicos.

El **Árbol de Decisión**, con una precisión de **0.7676**, siendo más sencillo y fácil de interpretar, demostró ser menos efectivo en comparación con los otros modelos. Este desempeño puede atribuirse a su tendencia a sobre ajustarse a los datos de entrenamiento, lo que limita su capacidad de generalización a nuevos datos. Por otro lado, el modelo de **Redes Neuronales**, con una precisión de **0.796** y una pérdida de **0.495**, mostró una ligera mejora para modelar relaciones complejas, pero aún así, su rendimiento fue inferior al de **RF**, lo cual sugiere que se podría mejorar con una optimización más exhaustiva de sus parámetros y construcción del entrenamiento.

El modelo de **Random Forest**, con una precisión de **0.8293**, sobresalió como el modelo más confiable para este conjunto de datos. Su capacidad para promediar los resultados de múltiples árboles de decisión reduce significativamente el riesgo de sobreajuste y mejora la generalización, permitiéndole capturar de manera efectiva la variabilidad y complejidad de las variables meteorológicas. Este rendimiento superior sugiere que el RF es particularmente adecuado para problemas de clasificación en contextos con datos complejos y de alta dimensionalidad.

Oportunidades Futuras de Investigación

Para futuras investigaciones, una primera área a explorar sería la optimización de hiperparámetros para cada uno de los modelos, especialmente las **Redes Neuronales**, ya que una búsqueda más exhaustiva podría mejorar significativamente su rendimiento. Además, se podría investigar la creación de nuevas características a partir de las variables existentes, lo cual podría ayudar a capturar mejor las relaciones subyacentes en los datos.

Otra oportunidad prometedora es la combinación de modelos híbridos que aprovechen las fortalezas de diferentes enfoques. Por ejemplo, utilizar un Random Forest para la selección de características y luego aplicar una Red Neuronal sobre estas características seleccionadas podría combinar la capacidad de generalización del Random Forest con la potencia de modelado de relaciones complejas de las Redes Neuronales.

Implementar técnicas de validación cruzada para asegurar que los modelos no están sobreajustados y que los resultados son robustos y generalizables es otra línea de investigación importante. Finalmente, investigar el impacto del cambio climático en las variables meteorológicas y su relación con la precipitación podría proporcionar conclusiones valiosas y mejorar aún más la precisión predictiva de los modelos desarrollados.

Agradecimientos

Queremos expresar nuestro agradecimiento a Cristóbal Alarcón Pérez por su valiosa ayuda en la creación de un script de web scraping en Python. Su apoyo nos permitió obtener datos de la Dirección Meteorológica de Chile de manera más eficiente y en mayor cantidad, lo cual fue fundamental para el desarrollo de nuestro proyecto.

Bibliografía

1. Basha, C. Z., Bhavana, N., Bhavya, P., & Sowmya, V. (2020, July). Rainfall prediction using machine learning & deep learning techniques. In 2020 international conference on electronics and sustainable communication systems (ICESC) (pp. 92-97). IEEE.
2. Parmar, A., Mistree, K., & Sompura, M. (2017, March). Machine learning techniques for rainfall prediction: A review. In International conference on innovations in information embedded and communication systems (Vol. 3).
3. Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, 7, 100204.