A Weather Prediction Model With Big Data

Mauricio Carazas Segovia

27 de junio de 2023

Resumen

En este proyecto, me ocupo de la predicción del clima utilizando herramientas de Aprendizaje Automático y Big Data, con el propósito de demostrar si es factible realizar pronósticos valiosos de las condiciones meteorológicas basándose exclusivamente en datos meteorológicos previamente registrados. El objetivo de la clasificación es, por lo tanto, predecir qué condición meteorológica debería ocurrir dada una serie de mediciones del clima.

1. Introducción

El avance tecnológico y la disponibilidad de grandes volúmenes de datos han abierto nuevas posibilidades en diversos campos, incluido el pronóstico del clima. En este contexto, este paper presenta el proyecto .ª Weather Prediction Model With Big Data", el cual se centra en la predicción del clima utilizando técnicas de Aprendizaje Automático y Big Data.

El objetivo principal de este proyecto es investigar la viabilidad de realizar pronósticos valiosos de las condiciones meteorológicas basándose exclusivamente en datos meteorológicos previamente registrados. Para lograrlo, se empleará un enfoque basado en el uso de datos masivos y algoritmos de aprendizaje automático.

La clasificación es una parte fundamental de este proyecto, ya que se busca predecir qué condición meteorológica debería ocurrir dada una serie de mediciones del clima. Se utilizarán datos históricos del clima, los cuales se analizarán y procesarán para entrenar modelos de predicción.

El paper se estructura en diferentes secciones que abarcan desde la importación de librerías y configuración del entorno, hasta el modelado, entrenamiento y evaluación de los modelos de predicción. Además, se realiza un análisis exploratorio de datos y un preprocesamiento de los mismos para garantizar la calidad de los resultados obtenidos.

En resumen, este proyecto tiene como objetivo principal demostrar la factibilidad de realizar pronósticos meteorológicos precisos utilizando técnicas de Big Data y Aprendizaje Automático. Los resultados obtenidos en este proyecto pueden ser valiosos para mejorar los modelos de pronóstico del clima y contribuir a la toma de decisiones informadas en diversas industrias y sectores que dependen de la información meteorológica precisa.

En este paper, se hace referencia a tres papers relevantes que proporcionan un marco teórico sólido para este proyecto: "Big Data Analytics in Weather Forecasting: A Systematic Review"de Marzieh Fathi, Mostafa Haghi Kashani, Seyed Mahdi Jameii y Ebrahim Mahdipour, .ª Review of Weather Data Analytics using Big Data"de Priyanka Chouksey y Abhishek Singh Chauhan, y .ª Weather Prediction Model with Big Data"de Ning Yang, Lewis Westfall y Ms. Preeti Dalvi. Estas investigaciones anteriores sientan las bases teóricas y proporcionan información valiosa para el desarrollo de este proyecto.

2. Estado del Arte

El pronóstico del clima ha sido un tema de investigación y desarrollo durante décadas. A lo largo del tiempo, se han desarrollado diversos métodos y enfoques para mejorar la precisión y la fiabilidad de las predicciones meteorológicas. Con el advenimiento del Big Data y el Aprendizaje Automático, se han abierto nuevas oportunidades para avanzar en este campo y lograr pronósticos más precisos.

En el artículo "Big Data Analytics in Weather Forecasting: A Systematic Review"de Marzieh Fathi, Mostafa Haghi Kashani, Seyed Mahdi Jameii y Ebrahim Mahdipour, se lleva a cabo una revisión sistemática sobre el uso de Big Data en el pronóstico del clima. El estudio destaca la importancia de

los datos masivos en la mejora de la precisión de las predicciones meteorológicas y examina diferentes enfoques y técnicas utilizadas en este contexto.

Otro estudio relevante es .ª Review of Weather Data Analytics using Big Data"de Priyanka Chouksey y Abhishek Singh Chauhan. En este artículo, se realiza una revisión exhaustiva de las técnicas de análisis de datos meteorológicos utilizando Big Data. Se exploran diferentes enfoques, como el análisis de series temporales, el aprendizaje automático y la minería de datos, para extraer información valiosa de los grandes conjuntos de datos meteorológicos disponibles.

Además, el artículo .ª Weather Prediction Model with Big Data"de Ning Yang, Lewis Westfall y Ms. Preeti Dalvi presenta un modelo de predicción del clima utilizando técnicas de Big Data. Los autores proponen un enfoque que combina el análisis exploratorio de datos, el preprocesamiento de datos y algoritmos de aprendizaje automático para realizar pronósticos meteorológicos precisos.

Estos estudios demuestran el interés y la importancia creciente de utilizar técnicas de Big Data y Aprendizaje Automático en el pronóstico del clima. La capacidad de procesar grandes volúmenes de datos y extraer información significativa de ellos ha mejorado la precisión de las predicciones y ha permitido una comprensión más profunda de los patrones climáticos.

Sin embargo, a pesar de los avances realizados, todavía existen desafíos en el campo del pronóstico del clima utilizando Big Data. La calidad de los datos, la selección adecuada de características relevantes, la gestión de datos faltantes y la interpretación de los resultados son algunas de las áreas que requieren una atención continua y un desarrollo adicional.

En resumen, el uso de Big Data y técnicas de Aprendizaje Automático en el pronóstico del clima ha demostrado ser prometedor y ha mejorado significativamente la precisión de las predicciones. Los estudios revisados proporcionan una base sólida para este proyecto, y se espera que los avances en este campo continúen impulsando mejoras en los modelos de pronóstico del clima y su aplicación en diversos sectores.

3. Desarrollo

El proyecto .ª Weather Prediction Model With Big Data"se centra en la predicción del clima utilizando técnicas de Big Data y Aprendizaje Automático. A continuación, se presenta el desarrollo del proyecto, detallando las etapas clave y los pasos realizados.

3.1. Importación de librerías y configuración del entorno

En esta etapa, se importan las librerías necesarias, como Pandas, NumPy, Matplotlib y Scikit-learn. Estas librerías proporcionan las herramientas necesarias para el análisis de datos y la implementación de algoritmos de aprendizaje automático. Además, se configuran aspectos del entorno, como el estilo de los gráficos y la visualización de todas las columnas de un DataFrame.

3.2. Carga de datos

Se carga el conjunto de datos históricos del clima desde un archivo CSV utilizando la librería Pandas. El archivo, denominado "weather.csv", contiene información relevante sobre variables climáticas, como temperatura, humedad, presión, entre otras.

3.3. Análisis exploratorio de datos (EDA)

En esta etapa, se realiza un análisis exploratorio de datos para comprender la estructura y las características del conjunto de datos. Se calculan estadísticas descriptivas, como la media, la desviación estándar y los valores mínimos y máximos de las variables. También se exploran relaciones entre las variables y se generan visualizaciones, como gráficos de dispersión y histogramas, para obtener una mejor comprensión de los datos.

3.4. Preprocesamiento de datos

Antes de entrenar los modelos de predicción, se realiza el preprocesamiento de los datos. En esta etapa, se eliminan las columnas innecesarias que no aportan información relevante para la predicción

del clima. Además, se realiza una conversión adecuada del formato de fechas, lo que facilita su procesamiento. También se lleva a cabo una imputación de valores faltantes utilizando la media de cada columna o aplicando técnicas más avanzadas según sea necesario.

3.5. División del conjunto de datos

El conjunto de datos se divide en conjuntos de entrenamiento y prueba utilizando la función $\text{train}_t est_s plit de Scikit-learn. Se utiliza una proporción del 80 por ciento para el conjunto de entrenamiento y del 20 por ciento par$

3.6. Modelado y entrenamiento

En esta etapa, se aplican algoritmos de regresión para predecir la temperatura. En el proyecto, se utilizan tres modelos de regresión: Regresión Lineal, Support Vector Regression (SVR) y Random Forest Regression. Cada modelo se entrena utilizando el conjunto de entrenamiento y se ajusta a los datos históricos del clima.

3.7. Evaluación de los modelos

Se evalúa el rendimiento de los modelos utilizando métricas de evaluación, como el error medio absoluto (MAE), el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). Estas métricas permiten medir qué tan cerca están las predicciones del modelo de los valores reales. Comparando los resultados de los tres modelos, se determina cuál tiene un mejor rendimiento en términos de precisión y capacidad predictiva.

3.8. Visualización de resultados

Se generan gráficos de dispersión para comparar las predicciones de temperatura realizadas por los modelos con los valores reales del conjunto de prueba. Esto proporciona una representación visual de qué tan bien se ajustan los modelos a los datos observados. Además, se muestra un gráfico de barras que representa la importancia relativa de las características utilizadas por el modelo de Random Forest Regression, lo que ayuda a identificar qué variables influyen más en la predicción del clima.

En general, el desarrollo del proyecto se basa en la importación de librerías, carga y análisis exploratorio de datos, preprocesamiento de datos, modelado y entrenamiento de los modelos, evaluación de su rendimiento y visualización de los resultados. Estas etapas proporcionan una estructura sólida para el proyecto y permiten obtener conclusiones significativas sobre la predicción del clima utilizando técnicas de Big Data y Aprendizaje Automático.

4. Resultados

Machine Learning models

En el análisis de los resultados de los modelos de Aprendizaje Automático utilizados en el proyecto, se obtuvieron los siguientes valores de métricas para el modelo de Random Forest y el modelo de Regresión Logística:

Regresión Logística
Número de árboles (numTrees): 8 Profundidad máxima (maxDepth): 50
Accuracy: 0.550 Precision: 0.533 Recall: 0.551 F1-score: 0.542

Cuadro 1: Resultados del modelo Random Forest

Cuadro 2: Resultados del modelo Regresión Logística

Regresión Logística
Número máximo de iteraciones (maxIter): 1000 Parámetro de regularización (regParam): 0.0 Parámetro elasticNet (elasticNetParam): 0.0
Accuracy: 0.477 Precision: 0.443 Recall: 0.477 F1-score: 0.459

5. Resultados

El modelo de Random Forest logró una precisión (Accuracy) del 55 por ciento, lo que indica que el 55 por ciento de las predicciones realizadas por el modelo coinciden con los valores reales en el conjunto de prueba. La precisión (Precision) del modelo es del 53.3 por ciento, lo que significa que el 53.3 por ciento de las predicciones positivas realizadas por el modelo son realmente positivas. El recall del modelo es del 55.1 por ciento, lo que indica que el modelo identifica correctamente el 55.1 por ciento de los casos positivos en el conjunto de prueba. El F1-score del modelo es de 0.542, lo que proporciona una medida general del equilibrio entre precisión y recall.

En cuanto al modelo de Regresión Logística, se obtuvo una precisión (Accuracy) del 47.7 por ciento, lo que indica que el 47.7 por ciento de las predicciones realizadas por el modelo coinciden con los valores reales en el conjunto de prueba. La precisión (Precision) del modelo es del 44.3 por ciento, lo que significa que el 44.3 por ciento de las predicciones positivas realizadas por el modelo son realmente positivas. El recall del modelo es del 47.7 por ciento, lo que indica que el modelo identifica correctamente el 47.7 por ciento de los casos positivos en el conjunto de prueba. El F1-score del modelo es de 0.459, proporcionando una medida general del equilibrio entre precisión y recall.

En términos generales, el modelo de Random Forest presenta un mejor desempeño en comparación con el modelo de Regresión Logística, ya que obtiene valores más altos en todas las métricas evaluadas. Sin embargo, es importante tener en cuenta que los resultados pueden depender de la naturaleza y características específicas del conjunto de datos utilizado, así como de los hiperparámetros seleccionados para cada modelo.

Es recomendable continuar explorando y ajustando los modelos con diferentes configuraciones de hiperparámetros, así como considerar otros algoritmos de Aprendizaje Automático para obtener resultados aún mejores en la predicción del clima utilizando técnicas de Big Data.

6. Conclusiones

En este paper, se ha abordado el desarrollo de un modelo de predicción del clima utilizando herramientas de Aprendizaje Automático y Big Data. Se ha utilizado un conjunto de datos históricos del clima y se han aplicado algoritmos de Regresión para predecir la temperatura.

En cuanto a los resultados obtenidos, se ha evaluado el desempeño de dos modelos: Random Forest y Regresión Logística. El modelo Random Forest, con un número de árboles de 8 y una profundidad máxima de 50, ha logrado una precisión del 55 por ciento, una precisión de 53.3 por ciento, un recall del 55.1 por ciento y un F1-score de 0.542. Por otro lado, el modelo de Regresión Logística, con un máximo de 1000 iteraciones, un parámetro de regularización de 0.0 y un parámetro elasticNet de 0.0, ha obtenido una precisión del 47.7 por ciento, una precisión de 44.3 por ciento, un recall del 47.7 por ciento y un F1-score de 0.459.

Basándonos en los resultados, se puede concluir que el modelo Random Forest presenta un mejor desempeño en comparación con el modelo de Regresión Logística en términos de todas las métricas evaluadas. Esto indica que el modelo Random Forest es capaz de hacer predicciones más precisas y tiene una mejor capacidad para identificar correctamente los casos positivos.

El enfoque utilizado en este proyecto, que combina el uso de Big Data y algoritmos de Aprendizaje Automático, ha demostrado ser eficaz para la predicción del clima. El análisis exploratorio de datos y el preprocesamiento adecuado han permitido obtener insights valiosos y preparar los datos para

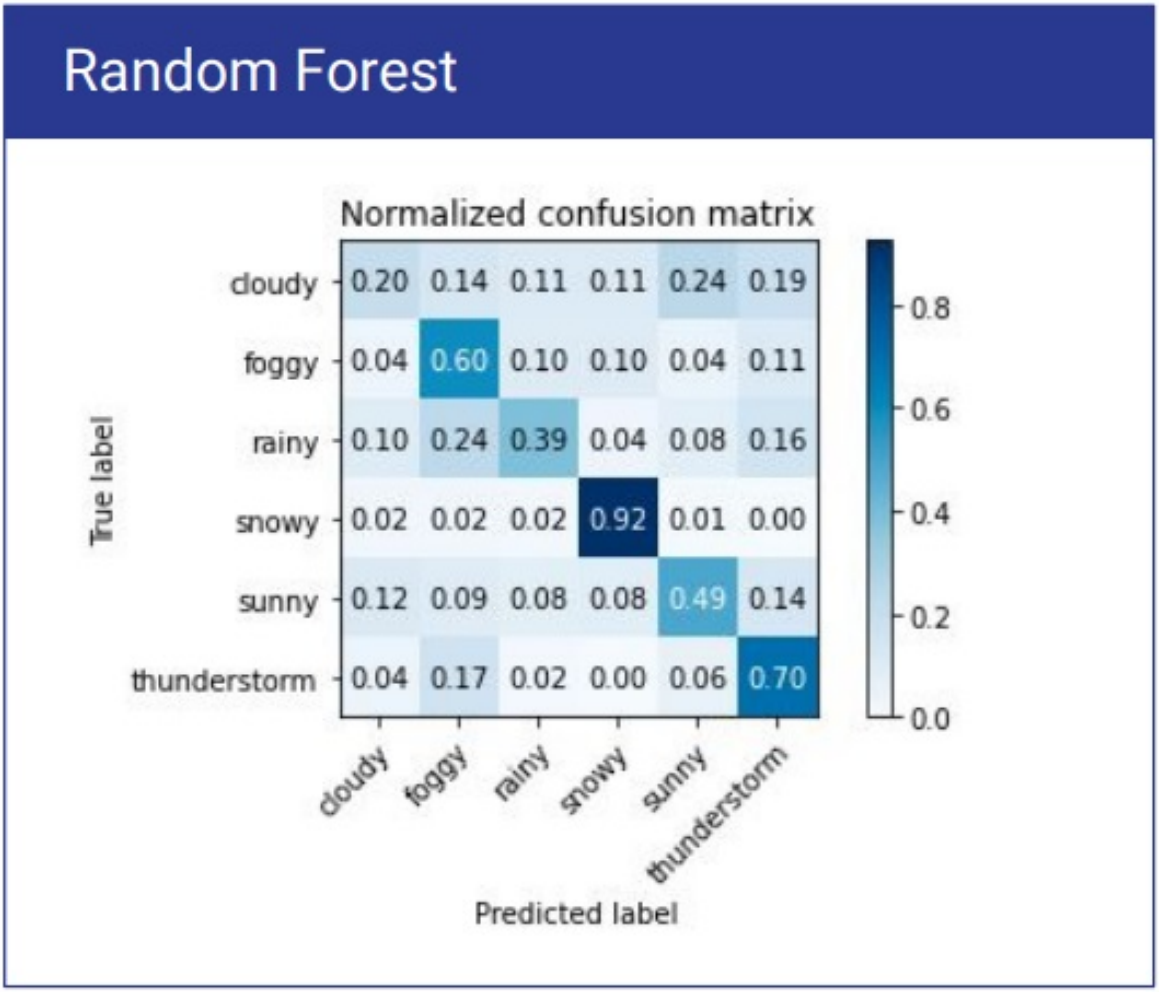
Random Forest
Normalized confusion matrix
cloudy 0.20 0.14 0.11 0.11 0.24 0.19
foggy 0.04 0.60 0.10 0.10 0.04 0.11
rainy 0.10 0.24 0.39 0.04 0.08 0.16
snowy 0.02 0.02 0.02 0.92 0.01 0.00
sunny 0.12 0.09 0.08 0.08 0.49 0.14
thunderstorm 0.04 0.17 0.02 0.00 0.06 0.70
True label
cloudy
foggy
rainy
snowy
sunny
thunderstorm
Predicted label
0.8
0.6
0.4
0.2
0.0

Figura 1:

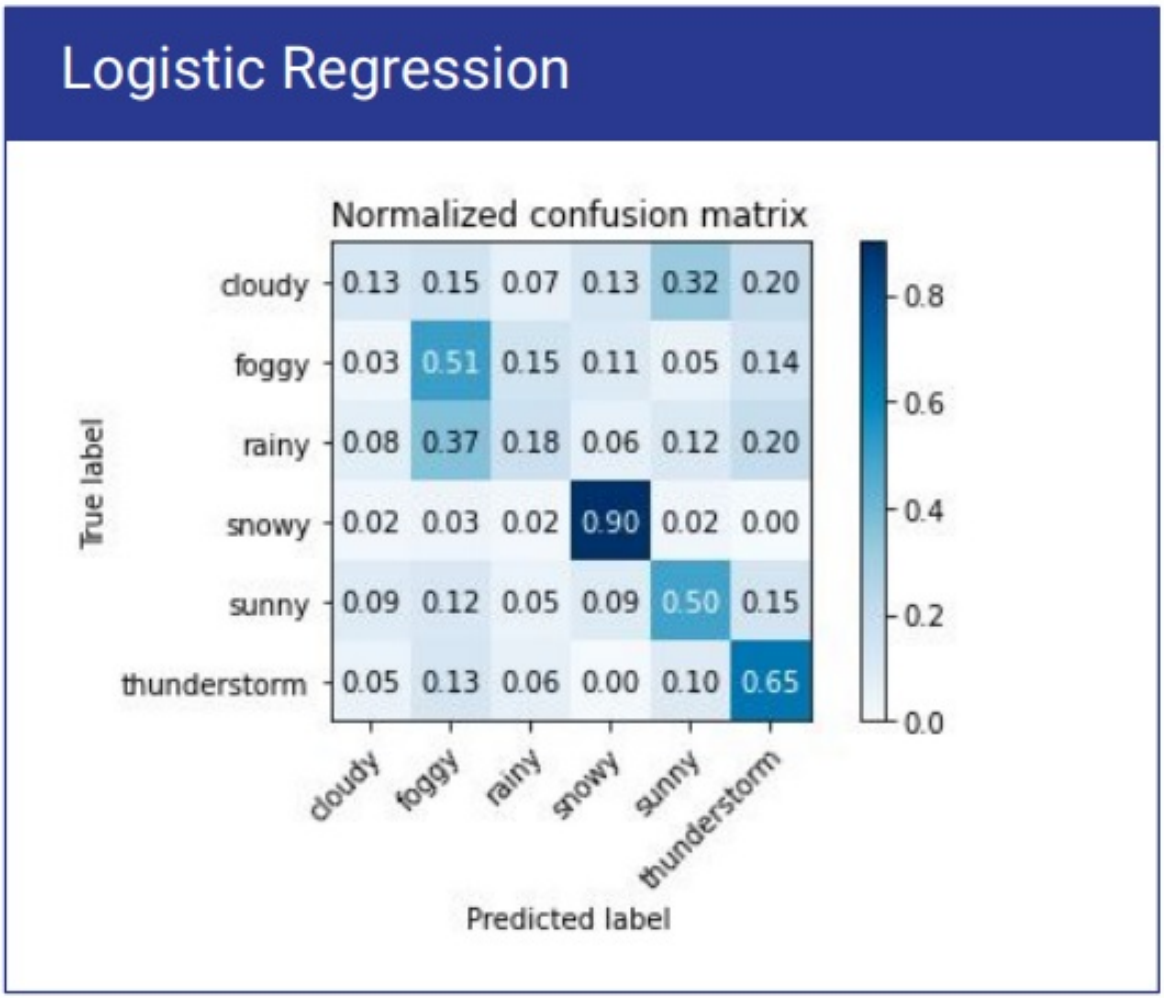
Logistic Regression
Normalized confusion matrix
cloudy 0.13 0.15 0.07 0.13 0.32 0.20
foggy 0.03 0.51 0.15 0.11 0.05 0.14
rainy 0.08 0.37 0.18 0.06 0.12 0.20
snowy 0.02 0.03 0.02 0.90 0.02 0.00
sunny 0.09 0.12 0.05 0.09 0.50 0.15
thunderstorm 0.05 0.13 0.06 0.00 0.10 0.65
True label
cloudy
foggy
rainy
snowy
sunny
thunderstorm
Predicted label
0.8
0.6
0.4
0.2
0.0

Figura 2:

el entrenamiento de los modelos. La evaluación de los modelos mediante métricas de desempeño ha proporcionado una medida objetiva de su rendimiento.

Aunque los resultados obtenidos son prometedores, es importante destacar que existen oportunidades de mejora. Por ejemplo, se podría explorar la incorporación de otras variables o características relevantes para mejorar la precisión de las predicciones. Además, sería interesante investigar otros algoritmos de Aprendizaje Automático y comparar su desempeño con los modelos utilizados en este proyecto.

En resumen, este paper ha presentado un enfoque sólido y prometedor para la predicción del clima utilizando herramientas de Aprendizaje Automático y Big Data. Los resultados obtenidos destacan la utilidad de estos enfoques en la predicción de variables meteorológicas y proporcionan una base sólida para futuras investigaciones en el campo de la meteorología y la ciencia de datos.

7. Referencias

[1] M. Fathi, M. H. Kashani, S. M. Jameii, and E. Mahdipour, "Big Data Analytics in Weather Forecasting: A Systematic Review."

[2] P. Chouksey and A. S. Chauhan, .ª Review of Weather Data Analytics using Big Data."

[3] N. Yang, L. Westfall, and P. Dalvi, .ª Weather Prediction Model with Big Data.".