CATCHING FIRE…
Before it Catches Us

A Wildfire Prediction Model

By Tina Salihu

Contents

INTRODUCTION

What is a Wildfire?

An uncontrolled fire that burns vegetation in natural areas like forests or grasslands.

What are the Risks Associated with Wildfires?

- Injury and loss of life
- Destruction of habitats and wildlife
- Economic loss

OBJECTIVE

How can we help to prevent these risks?

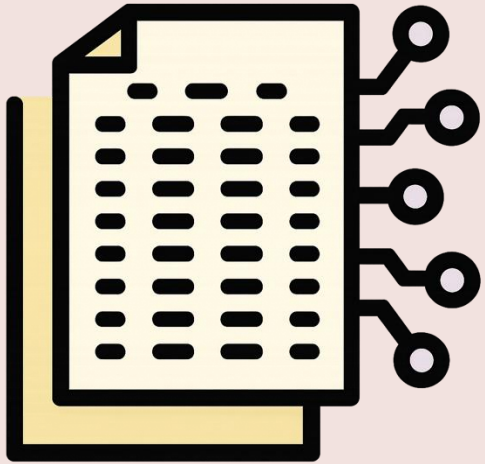
Build a predictive model!

OBJECTIVE: Build a classification model that takes information about environmental factors and calculates the probability of a wildfire occurring.

This can be used to send warning messages to help with planning and risk-management.

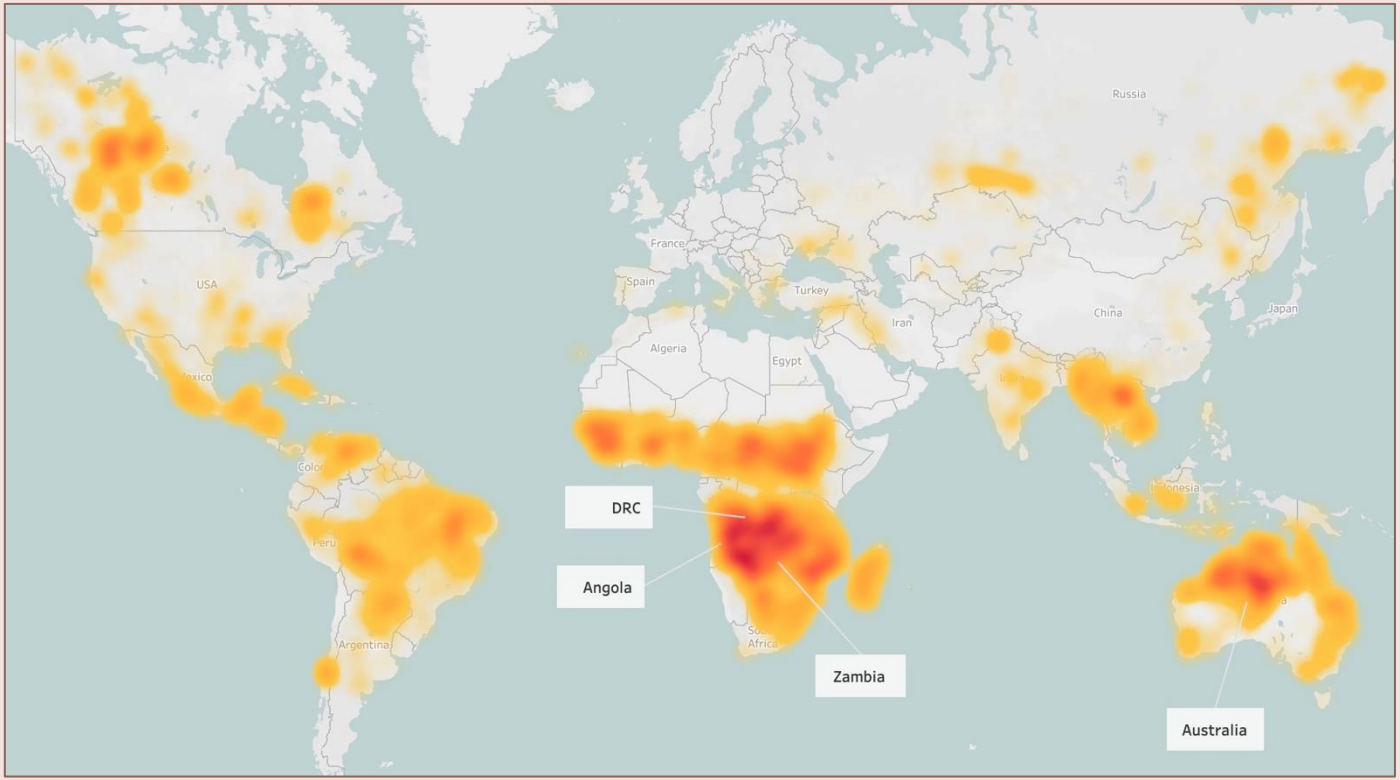
EXPLORATORY DATA ANALYSIS

The dataset chosen was a CSV file from **Kaggle**, but was originally scraped from **NASA** and **Open-Meteo**:



- 15 Features
  - Includes temperature, wind speed, humidity.
  - *Represent measurements over a one-hour period.*
- Target Column: '**Occurred**'
  - Binary classification, i.e. 0 = No Wildfire, 1 = Wildfire

EXPLORATORY DATA ANALYSIS

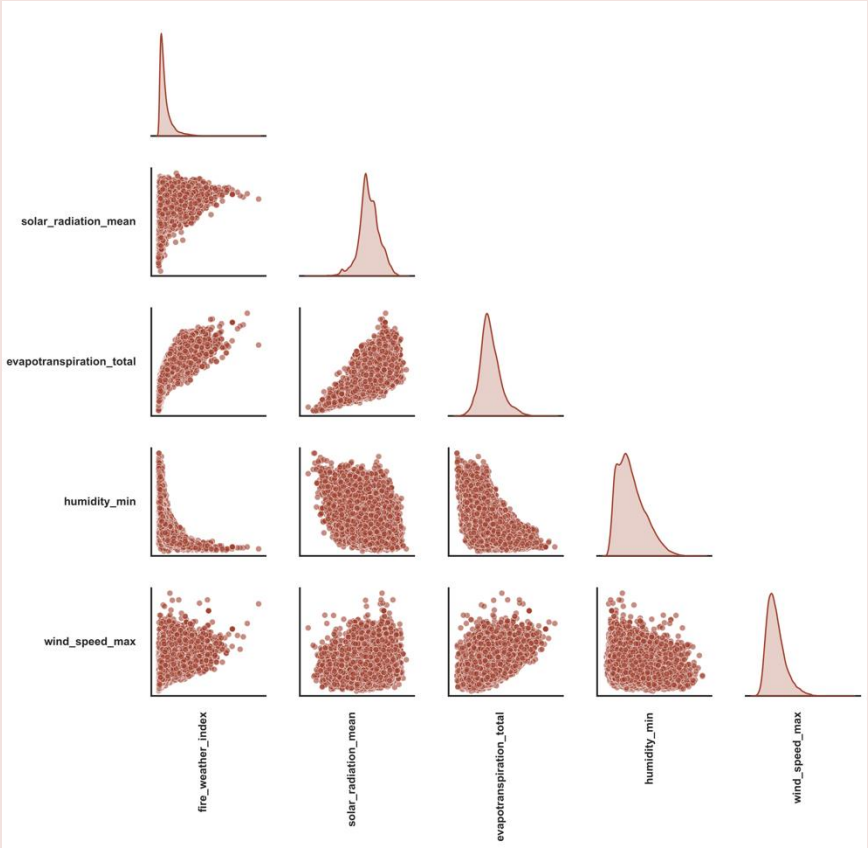


Wildfire Hotspots:

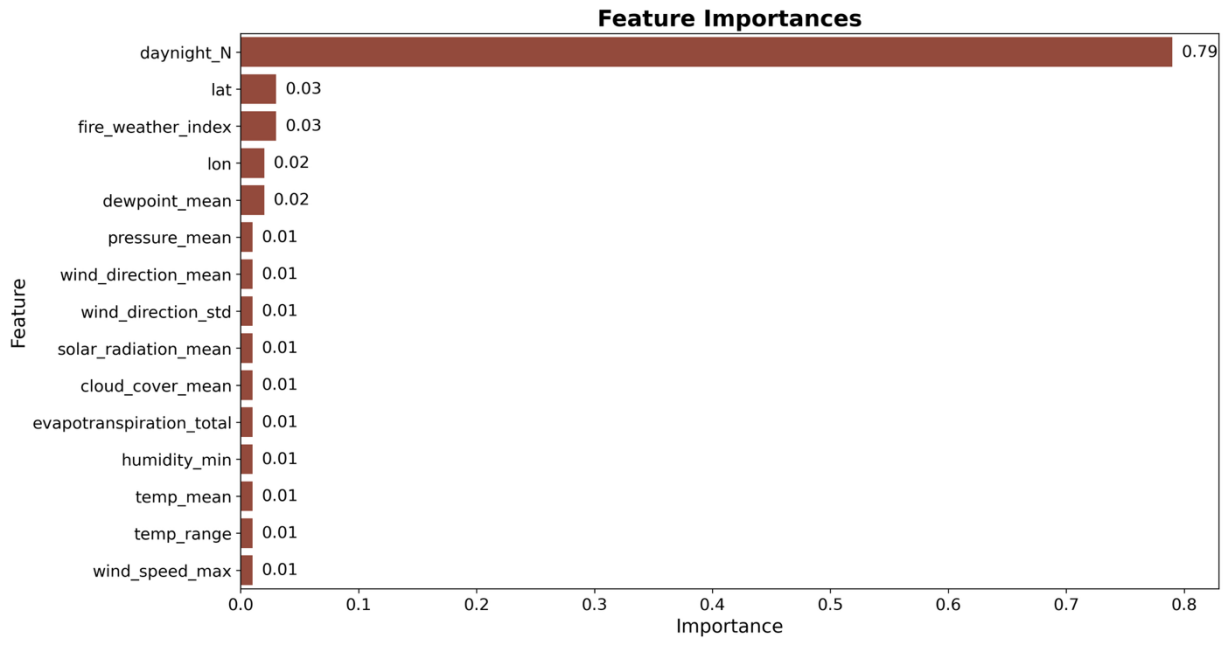
- Congo, Angola, Zambia
- Australia
- South America

EXPLORATORY DATA ANALYSIS

- Some columns were found to have **non-linear relationships**.
- So **logarithmic transformations** were applied to these columns to improve *model performance*.



DATA EXPLORATION



- The '**daynight_N**' column – *a binary column indicating whether it is day or night* – had the highest feature importance.

- Columns with a **low feature importance** were removed initially, but this *negatively* affected the model's performance, so they were added back in.

THE MODEL

The model chosen was **XGBoost** due to its better performance over Logistic Regression and Random Forest.

XGBoost is a decision-tree model that builds many simple trees one after another, each one correcting the errors of the previous ones.

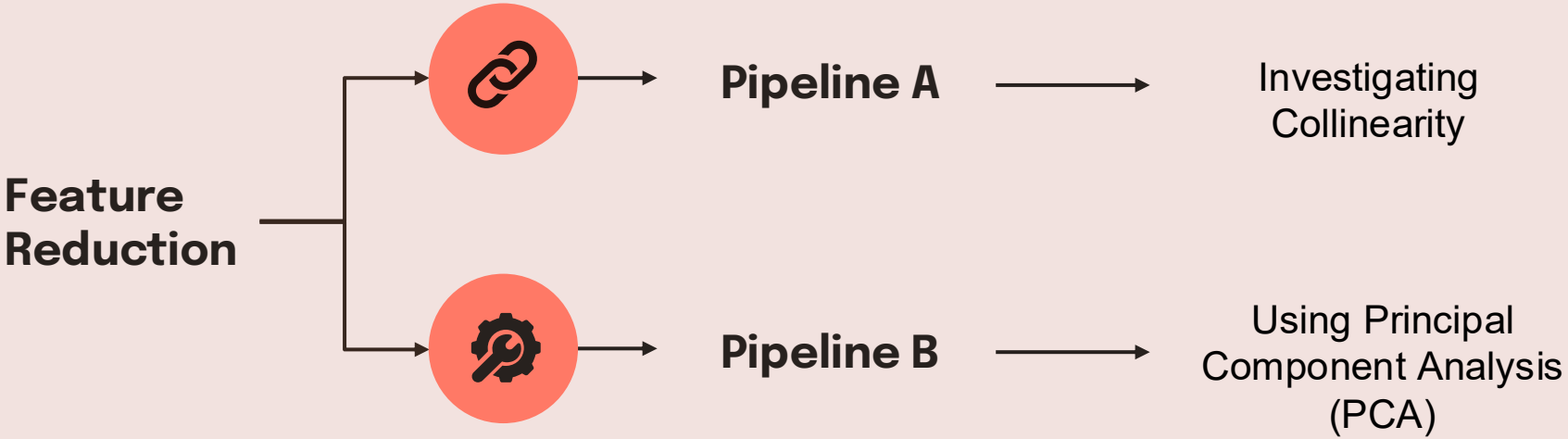
Advantages:

- Faster
- More accurate
- Better at handling messy data

THE MODEL – Feature Reduction

The wildfire dataset has 17 columns and over 100,000 rows.

Two pipelines were explored to reduce the number of features:



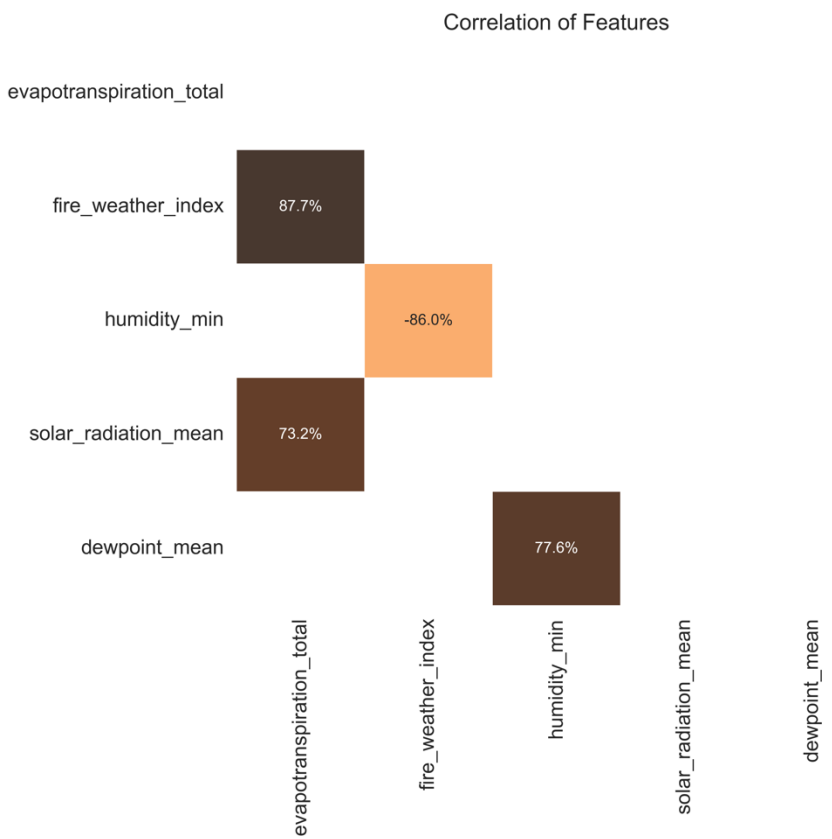
Pipeline A

Collinearity occurs when two features are very closely related.

1. Correlations were calculated between features

2. Variance-Inflation-Factor (VIF) scores were calculated for features with high correlations

3. Columns with high VIFs were dropped from the model.



Pipeline B

Principal Component Analysis (PCA) reduces the number of columns while keeping the important information.

How it Works:

1. Analyses features to find patterns in the data
2. Combines these into a smaller set of features
3. Keeps 95% of the variance in the data.

PCA performed better, so Pipeline B was chosen for the final model.

THE MODEL – Evaluation

Recall measures the proportion of actual wildfires that the model correctly identifies.

High recall means the model misses very few real wildfires. Since our goal is to support **risk management**, we prioritise catching as many real wildfires as possible.

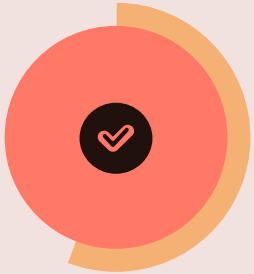
To achieve this, the model's **classification threshold** was set to **0.4** – any predicted probability of 40% or higher counts as a wildfire.

Results

94%

Recall

58%



Precision

71%

F1

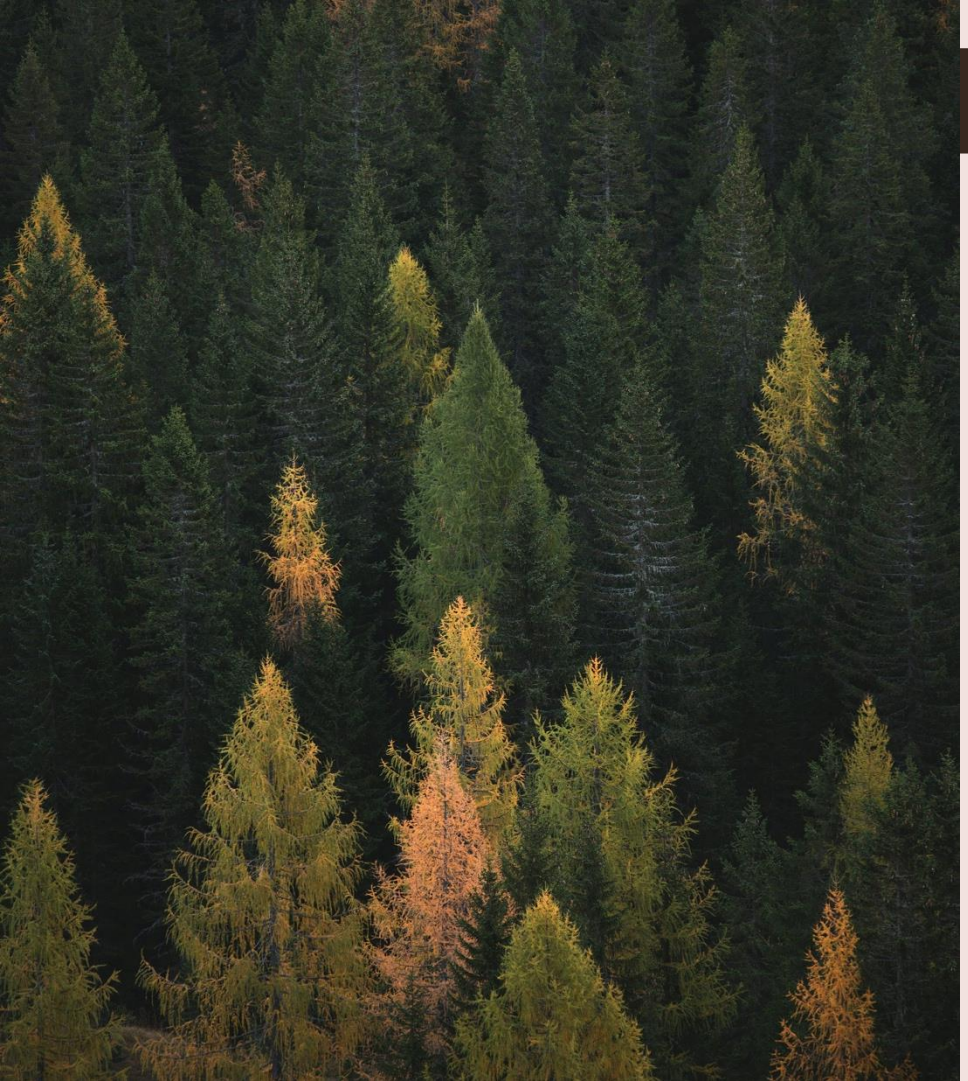
DEMONSTRATION

Future Proofing

01 — Incorporate multiple data sources with more relevant features

02 — Extend the model to predict fire *intensity*

03 — Extend the model to predict fire *spread*



CONCLUSION

The objective of predicting wildfire occurrence from environmental factors was achieved using an **XGBoost model** with a strong **94%** recall.

Day/night emerged as the most influential feature, demonstrating the model's potential for effective early warning and risk-management support.

Any Questions?