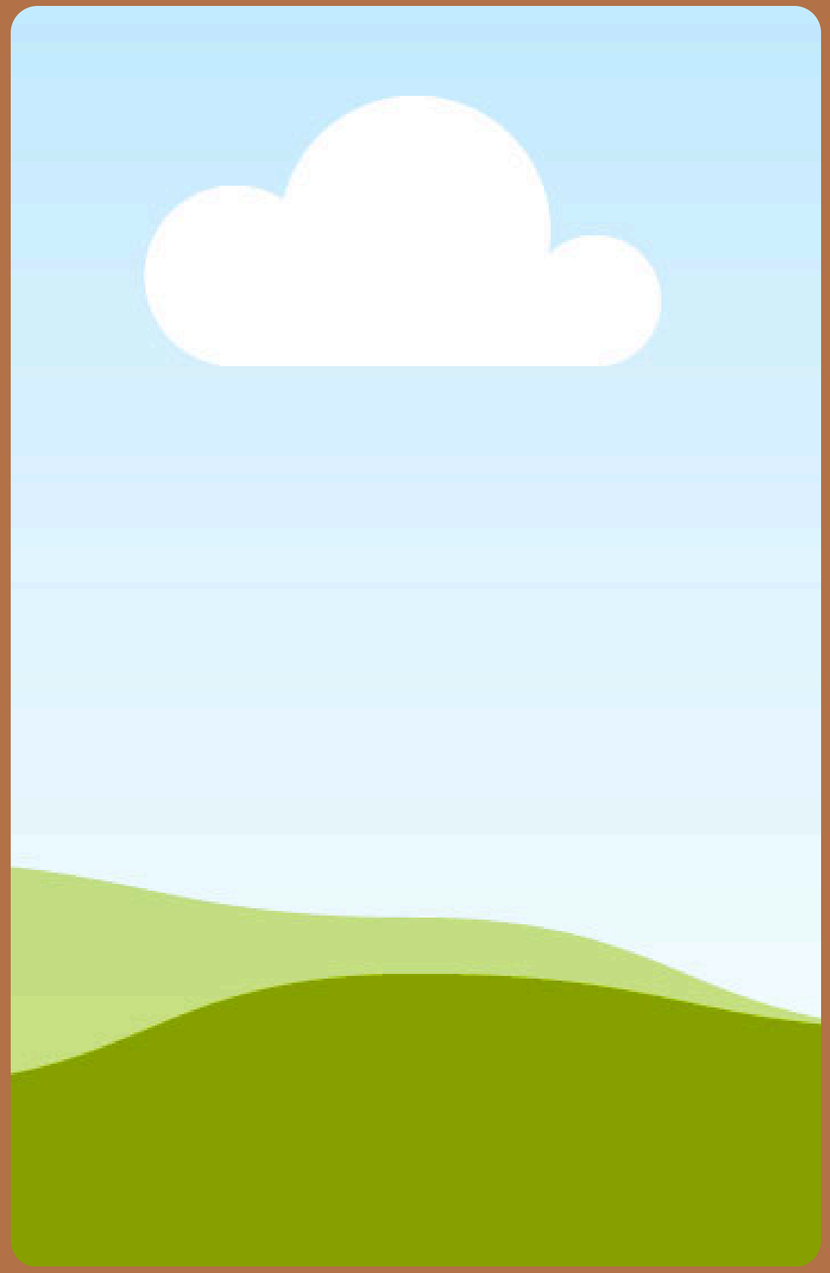
Model to Predict Coffee Disease Risk For Proactive Farm Management



Team members

- KIGEN TUWEI

CATHERINE KAINO

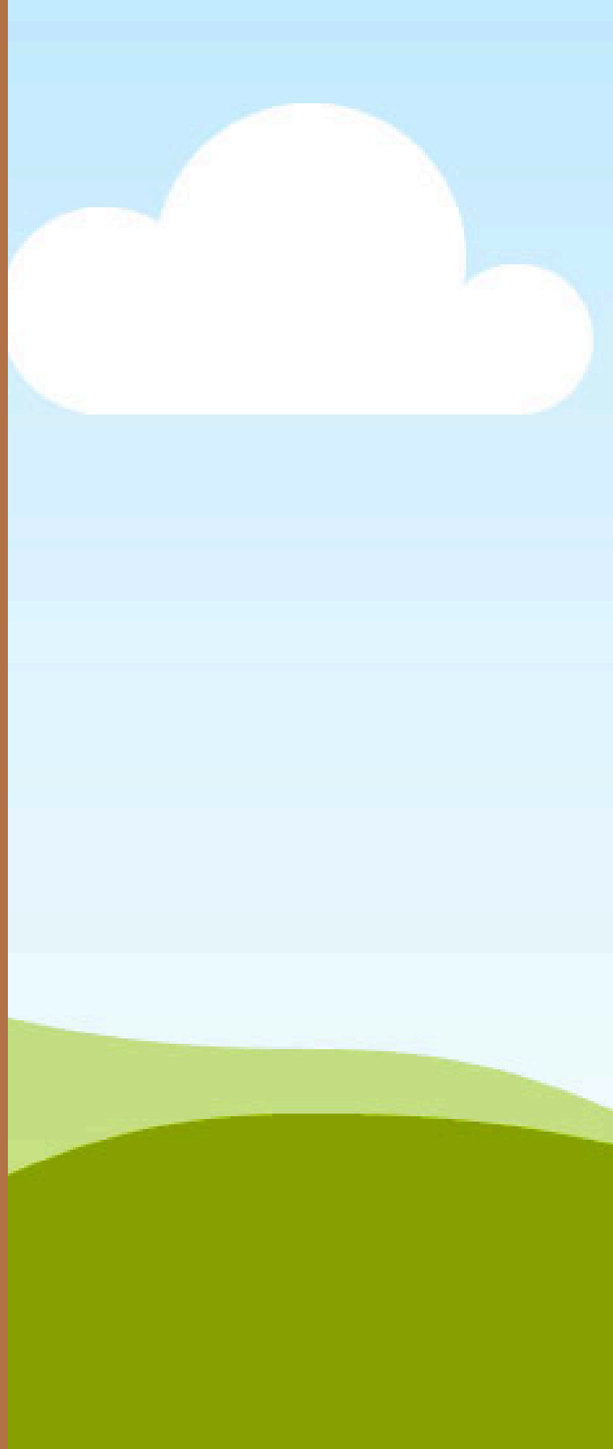
JUNE MASOLO

JORAM MUGESA

- KENNEDY OMORO

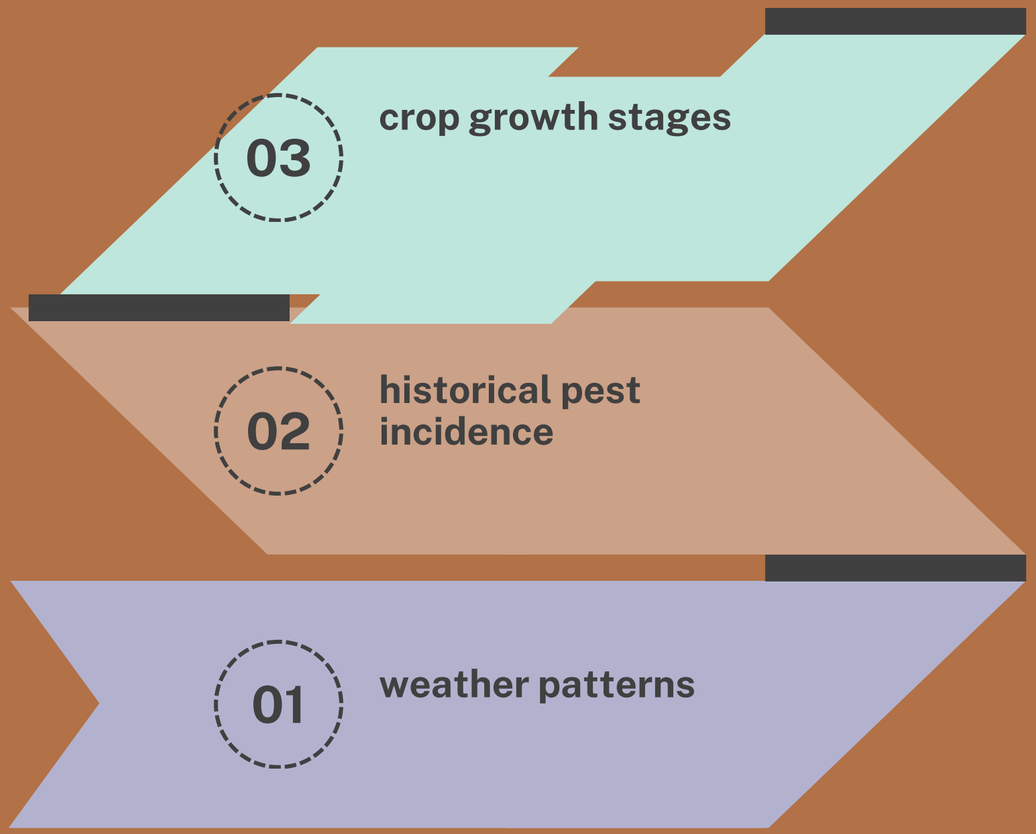
Problem Background

Coffee production in Kenya faces growing risk from unpredictable diseases like Coffee Leaf Rust, which can cut smallholder yields by up to 70% and cause financial instability. Current responses are mostly reactive, relying on late symptom detection or costly, indiscriminate fungicide use, resulting in economic losses and environmental harm. There is a major need for an early-warning system that uses environmental data to enable timely, proactive disease management.



Objectives

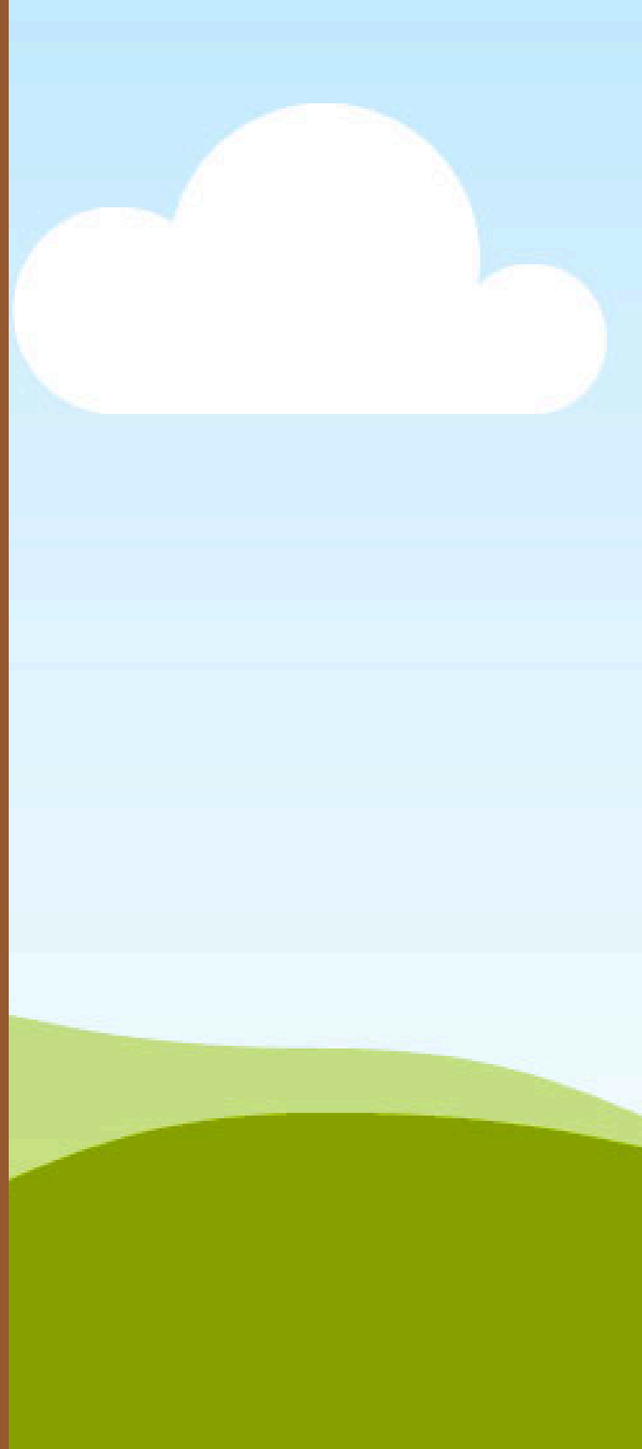
- This project entails building a supervised machine learning model to predict the risk level of coffee leaf rust disease outbreaks. The objective is to classify upcoming risk as Low, Medium, or High based on environmental and historical data, enabling farmers to apply fungicides or pesticides proactively and only when necessary.



Dataset Used

Weather Data: Historical and forecast meteorological data (temperature, humidity, rainfall) from the NASA POWER API, specifically tailored for agromodeling. • Link: https://power.larc.nasa.gov/

Extracting Data from NASA POWER API (The model will use Nasa Power dataset from 01-01-2010 to 31-12-2020 focusing on Coffee plantation Kenya in Nyeri area(major coffee zone)).



Methodologies

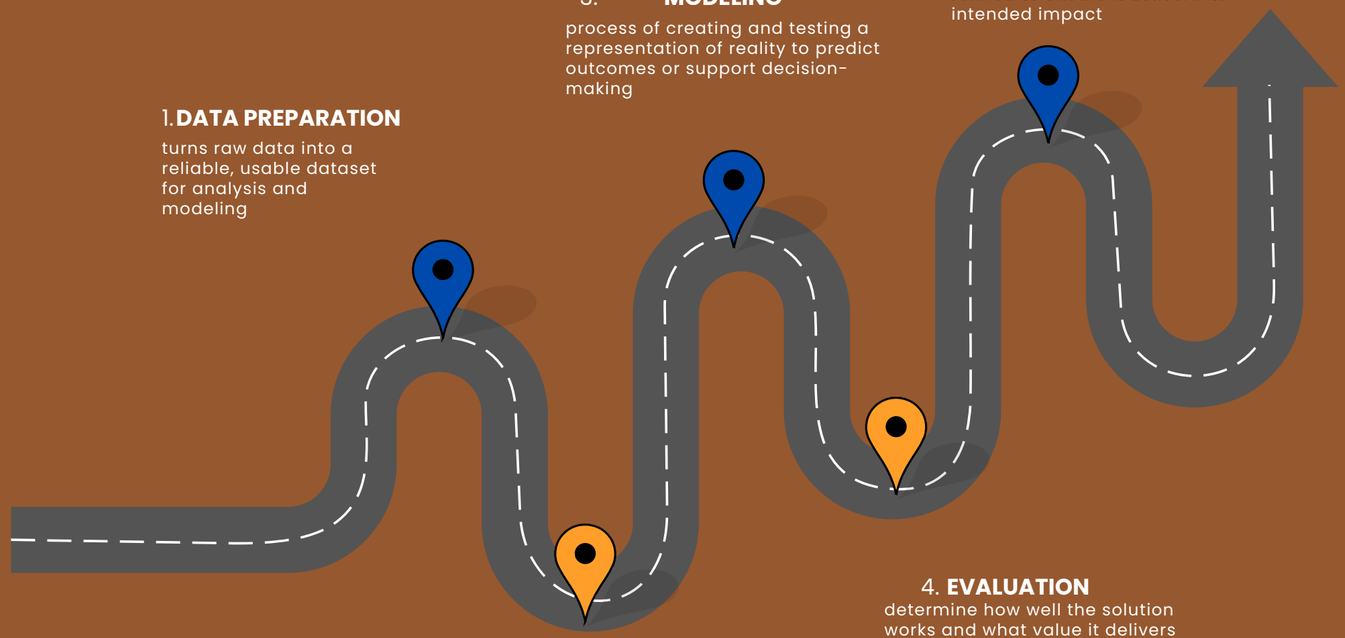
1. **DATA PREPARATION**
turns raw data into a reliable, usable dataset for analysis and modeling

2. **EXPLORATORY DATA ANALYSIS (EDA)**
make sense of the data and uncover insights before modeling

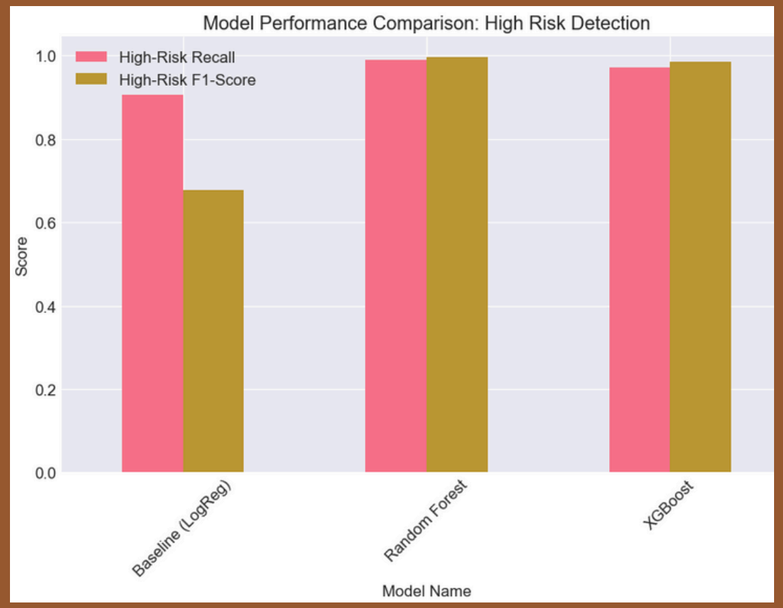
3. **MODELING**
process of creating and testing a representation of reality to predict outcomes or support decision-making

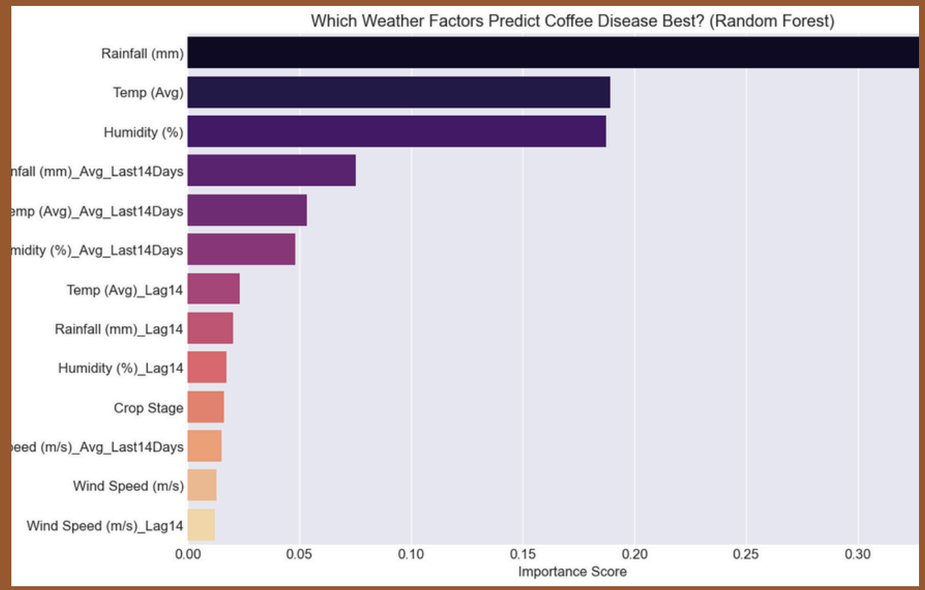
4. **EVALUATION**
determine how well the solution works and what value it delivers

5. **DEPLOYMENT**
A developed solution is rolled out, used in practice, and refined to ensure it delivers its intended impact



Evaluation Results



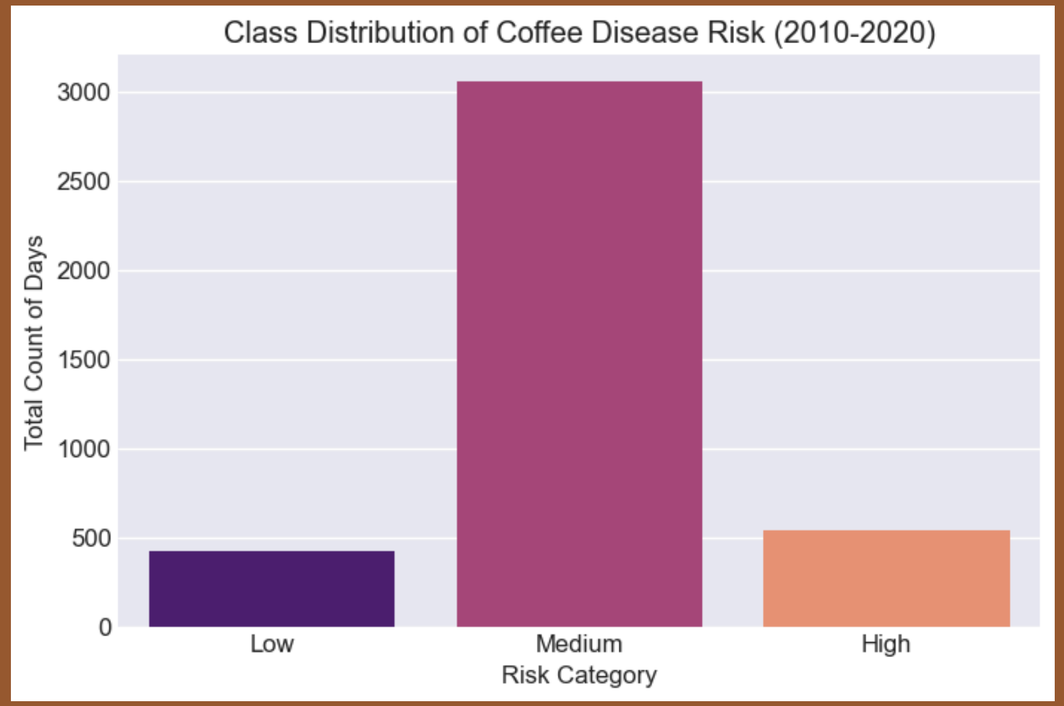


The Key Finding:
The baseline logistic regression model provides moderate accuracy but lacks reliability due to low precision, leading to many false outbreak warnings despite catching most real cases. In contrast, Random Forest and XGBoost perform exceptionally well, achieving near-perfect accuracy and precision. Random Forest is best at detecting nearly all high-risk outbreaks, while XGBoost delivers the highest overall accuracy across all classes.

The Key Finding: Central Tendency and Consensus

-
-

Key EDA Finding

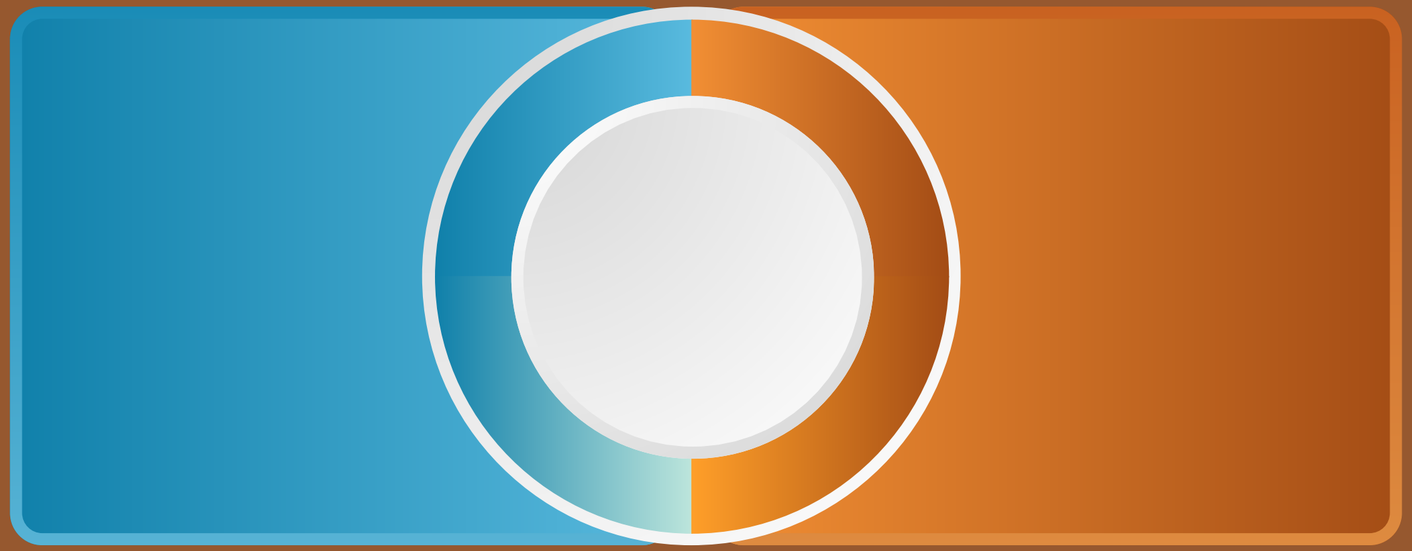


The data shows a moderate class imbalance, with far more “Medium” risk days than “High” risk days. To address the business impact of missing high-risk events, the model will use balanced class weights, ensuring that failing to detect a high-risk day is treated as significantly more costly than misclassifying a medium-risk day.

Your paragraph text

Modeling

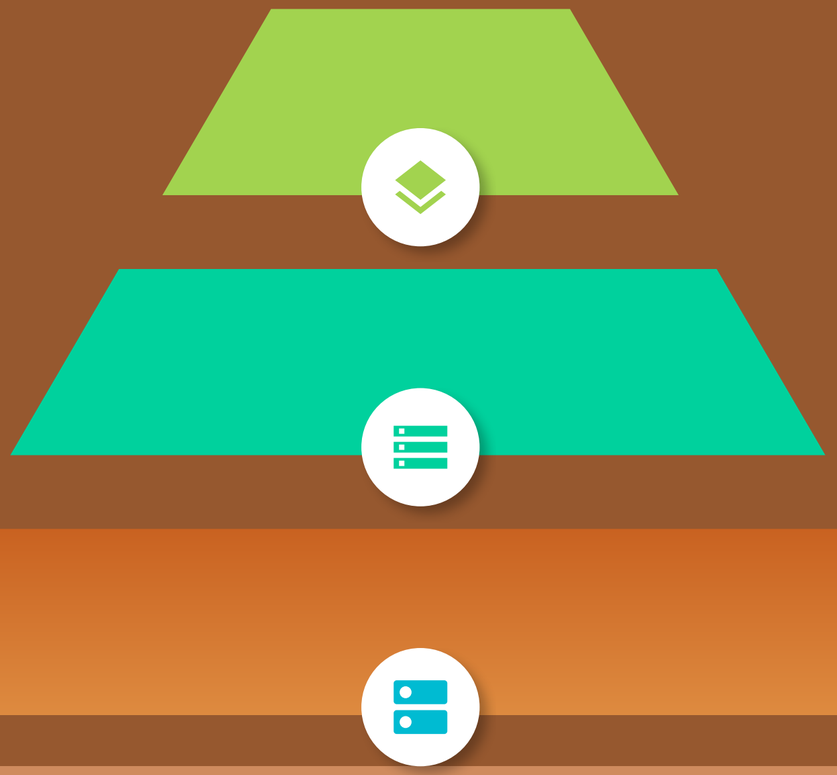
entails building three models(Logistic regression, Random Forest and XGBoost Classifier) using ML pipelines.



create a loop to run all three models. This allows us to see the Baseline (Logistic Regression) vs. the Ensemble models (Random Forest and XGBoost) side-by-side.

Your paragraph text

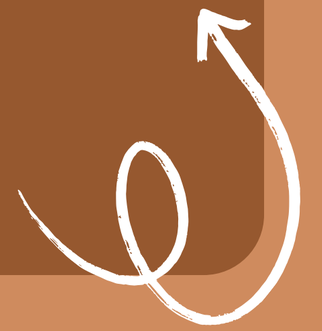
Model



M

•

Overall Interpretation:



Thank You!

We appreciate your questions and engagement