Wildfire Seasonal Feature importance

Problem Statement

The Challenge of Wildfire Prediction

Wildfires cause extensive damage to ecosystems, property, and human lives.

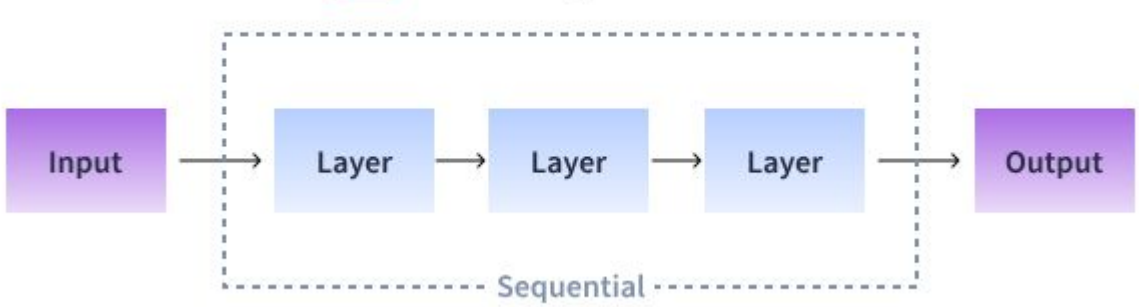
Predicting them is crucial for timely intervention and management.

Factors influencing wildfires include weather conditions, human activities, and natural causes.

Machine learning offers a robust approach to model and predict these events based on historical data.

Algorithm Discussion

Mia - Neural Networks

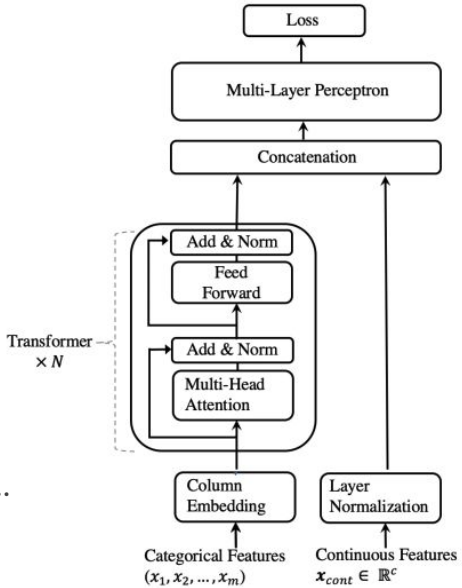


Tabular datasets are the last “unconquered castle” for deep learning, with traditional ML methods like Gradient-Boosted Decision Trees still performing strongly even against recent specialized neural architectures.

– A review paper for deep learning model with tabular datasets

Approach

1. Baseline model
   - Input + 2 hidden layers + output
   - Preliminary result
2. Increase depth of hidden layers based on baseline model
   - More units and more layers
   - Added Batch Normalization, Dropout
   - Improved result compared to baseline model but not much
3. Try different model architectures specific for tabular data
   - ~~TabTransformer~~ specific for categorical data while ours are numerical
4. Fine-tune NN model
   - Keras Tuner library
     i. Regularization: L1L2, dropout rate, batch normalization, early stop...
     ii. Layer units, count, activation functions...
     iii. Epoch, batch size, learning rate...



Feature Importance Analysis

Permutation feature importance

1. Calculate the error between prediction and actual labels as baseline error.
2. For each input feature, shuffle all test records of it and maintain the correct order for other features.
3. Predict with shuffled test data and measure the error after the shuffling.
4. Sort the error from large to small, then we get the order of each feature's importance.

The idea behind this method is, the more important a feature is, the more impact it has on the result. Shuffling such features will break the link between features and labels, thus causing error increasing.

Huizi - Random Forest

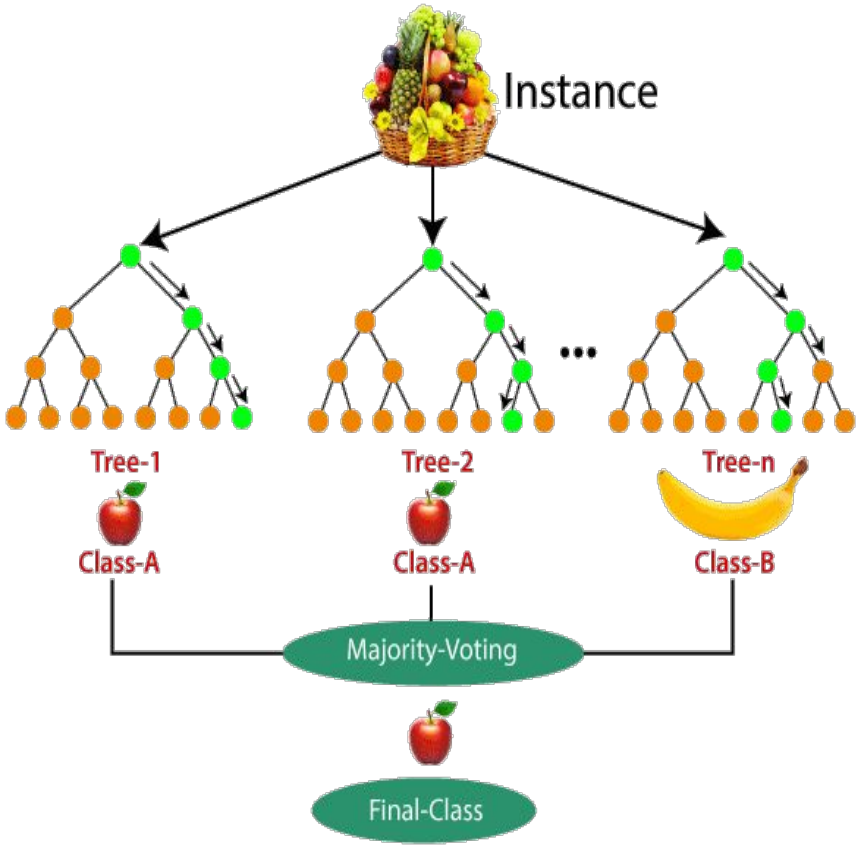
Ensemble learning method that combines multiple decision trees.

Offers higher accuracy through bagging and feature randomness.

Handles large data sets with higher dimensionality.

Can model nonlinear relationships.

Produces the mode (classification) or mean (regression) prediction of individual trees for unseen data.



Reference: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

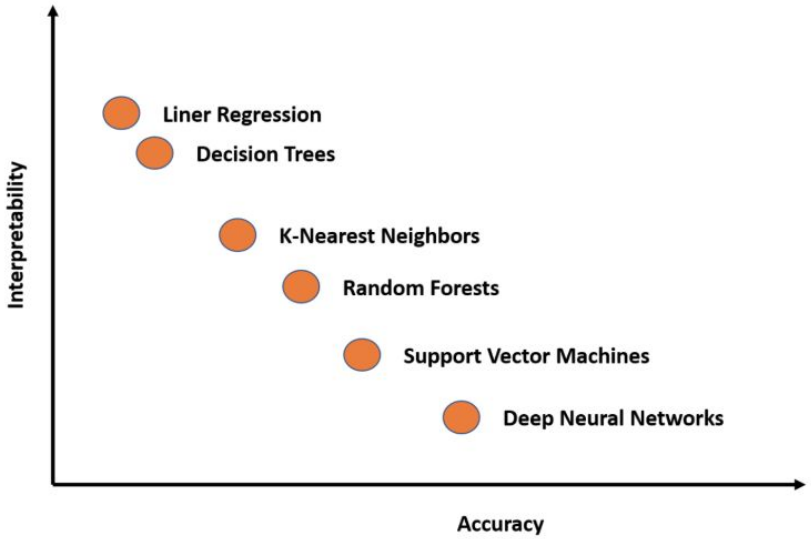
Why Choose Random Forest?

Handles missing values and maintains accuracy even when a large proportion of data is missing.

Provides a feature importance estimate, useful for understanding influential factors.

Effective in cases where model interpretability is crucial.

Resistant to overfitting due to averaging of multiple decision trees.



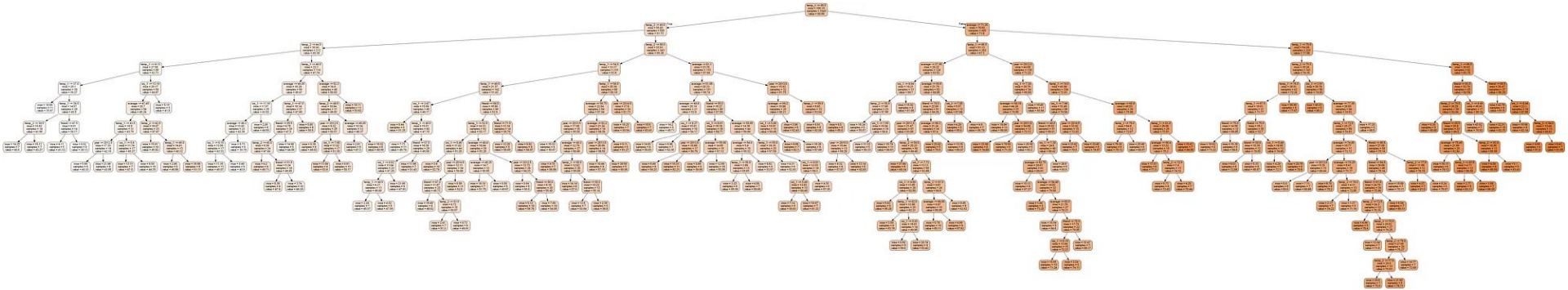
Reference: https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/

Pre-processing and Data Preparation

Explanation of data splitting.

Importance of balanced datasets.

Feature selection and engineering.



Reference: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

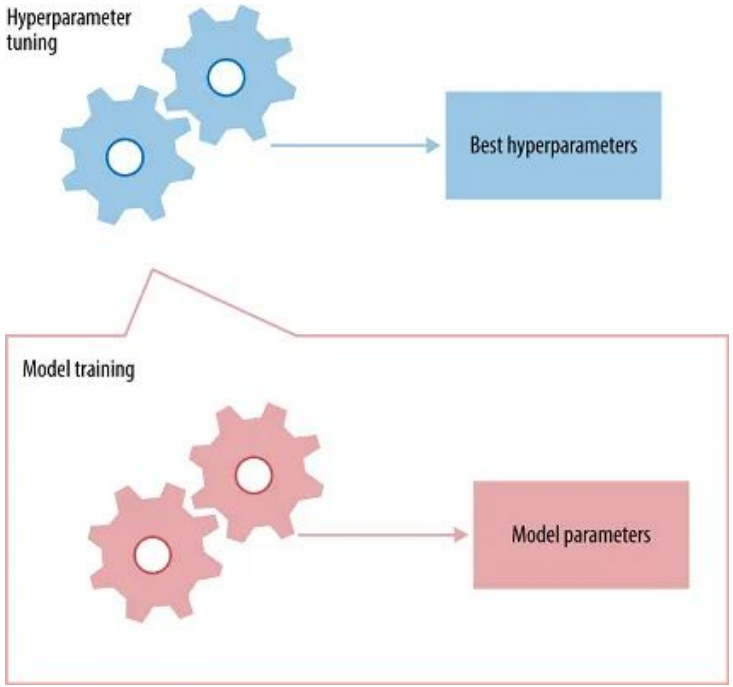
Hyperparameter Tuning

Importance of tuning for model optimization.

Explanation of overfitting.

Parameters tuned: n_estimators', 'min_samples_split', 'min_samples_leaf', 'max_depth', bootstrap'

Method chosen for tuning: Random Search Cross Validation in Scikit-Learn



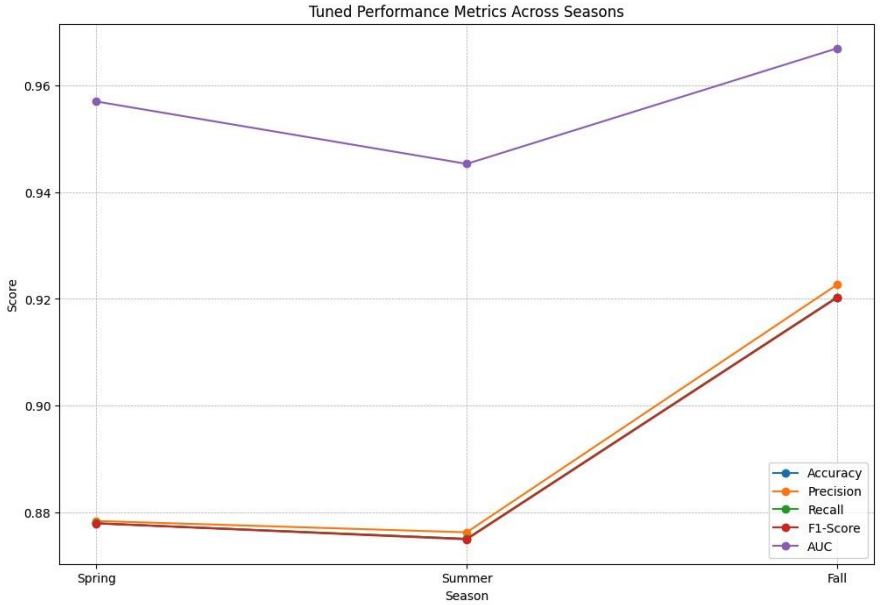
Reference: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Performance Metrics

Importance of evaluation metrics.

Introduction to Accuracy, Precision, Recall, F1-Score, and AUC.

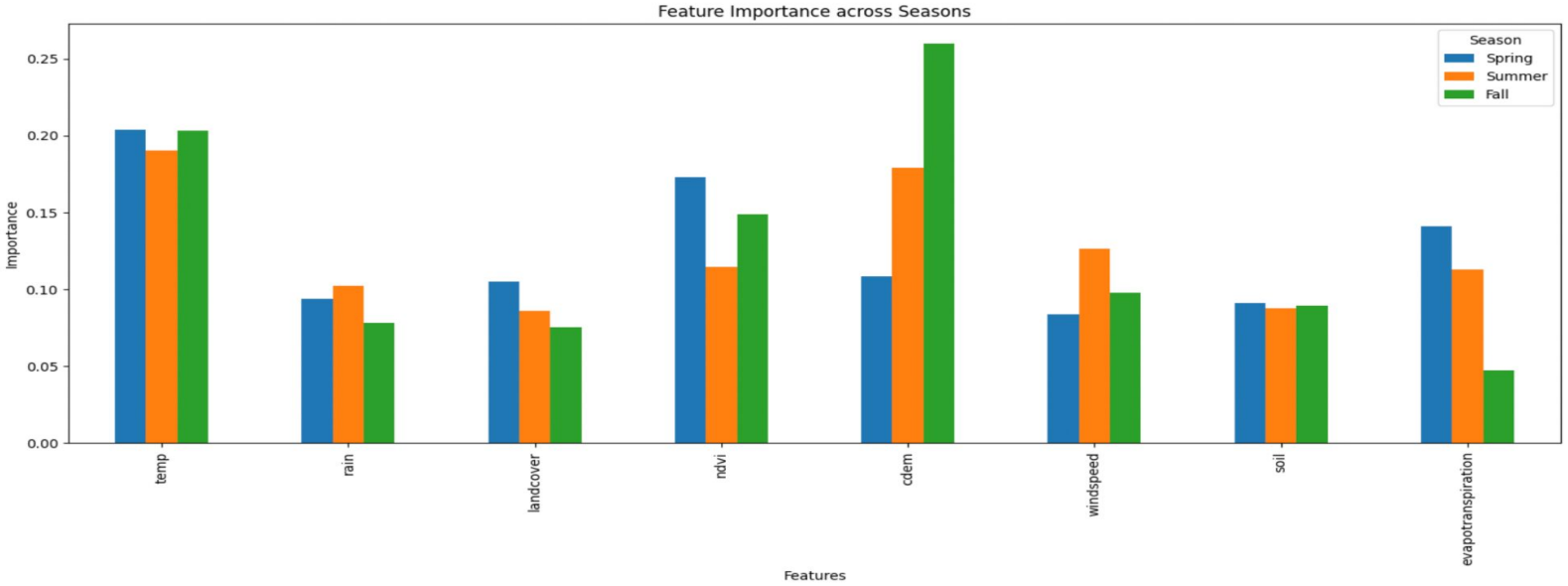
Comparison of metrics before and after tuning.



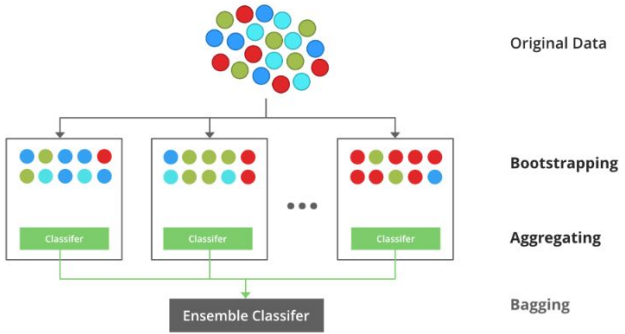
Feature Importance

Why understanding feature importance is crucial.

Graphical representation of features and their importance across seasons.

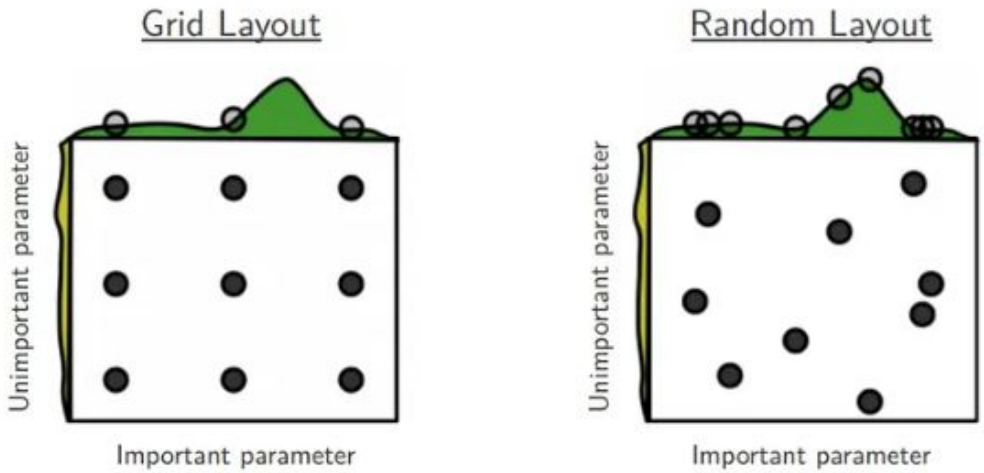


Approach: eXtreme Gradient Boosting (XGBOOST)



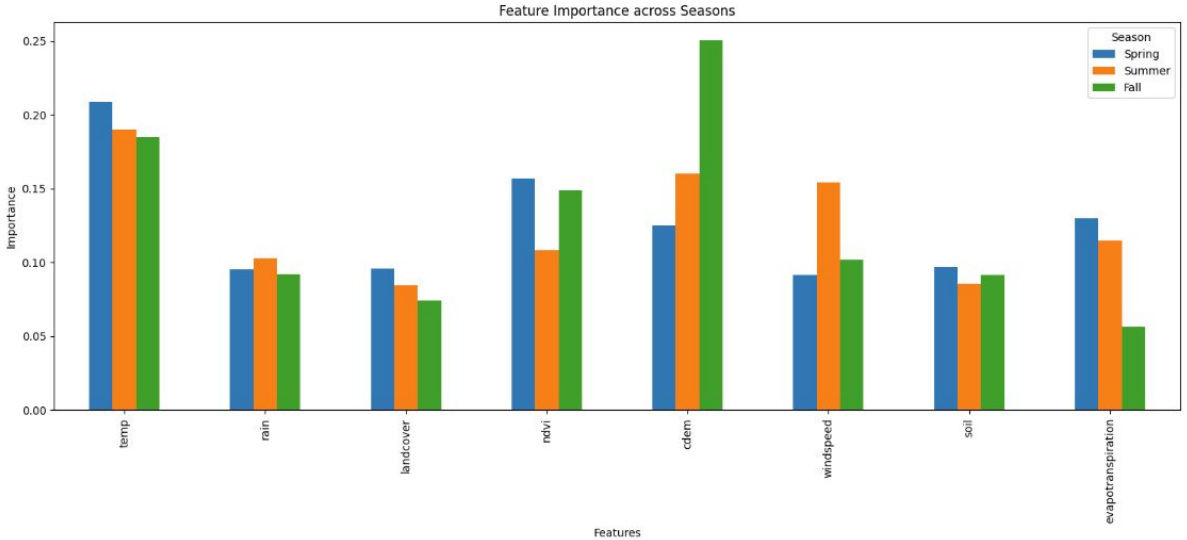
- XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) method.
- XGBoost was developed by Tianqi Chen and Carlos Guestrin, and it has gained prominence for its exceptional performance and wide range of applications in machine learning.

Hyperparameter Tuning: Randomized Search



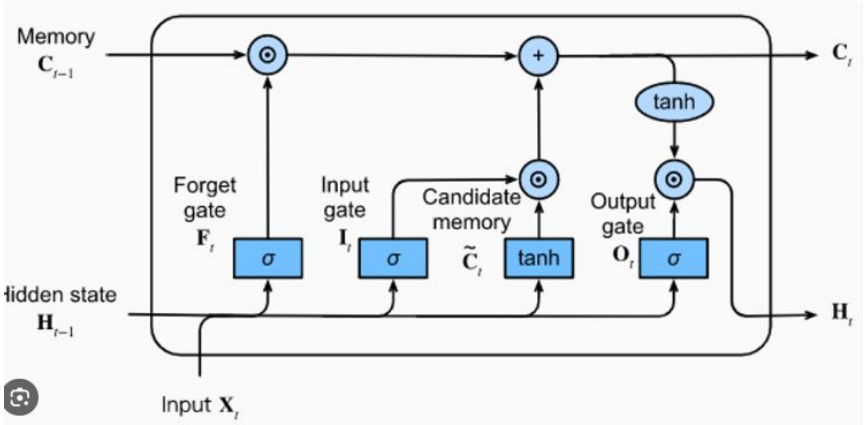
- More Efficient way for parameter search
- With grid search, nine trials only test three distinct places. With random search, all nine trails explore distinct values.

XGBOOST Feature Importance



- Use model.feature_importances_ in XGBOOST library
- Seasonal Variations: Investigate the changing importance of predictors across different seasons

Other ML Approaches



- We also experiment with **KNN, Kernel SVM, LSTM** for Wildfire prediction, and feature importance Analysis
- Result are not as good as Tree-based methods for now (~80 vs ~90 percentage in Accuracy, Recall)
- Next step is to deploy models from **Hugging face** using transformer for table data for further exploration and feature importance analysis

Simulation Plan

Dataset

The fire season defined by BC officials ranges from April to November. So we define the temporal scope at the scale of season as: Spring (3~5), Summer (6~8), Fall (9~11).

Given the temporal overlap of input maps available on Google Earth Engine, we choose training data from 2018-2020 and test data from 2021.

We stratified-sampled 800 points in each month for training and test.

In summary:

- Training dataset contains 7200 points for each season in three years.
- Test dataset contains 800 points for each season in 2021.

With stratified sample we make sure the dataset is balanced.

Sustainability

Sustainability in Wildfire Management

Using machine learning for prediction can lead to timely interventions, reducing the extent of damage.

Efficient resource allocation: Predictions can guide where resources (like firefighters or equipment) should be directed.

Long-term benefits: Understanding and mitigating wildfire risks can lead to healthier ecosystems and reduced economic losses.

Ethics in Machine Learning Predictions

Ensuring data privacy: Protecting the information of regions and individuals.

Avoiding bias: Ensuring the model doesn't unintentionally favor any specific group or region.

Transparency: Open discussions about how the model makes predictions and potential limitations.

Conclusion