# Predicting Wildfire Occurrence

## A Comparative Machine Learning Study Using Climate Indicators

SIMBANEGAVI SIMBARASHE
JEONGWON YOO
NITHIN RAVINDRA REDDY

# Table of contents

# 01

# Introduction

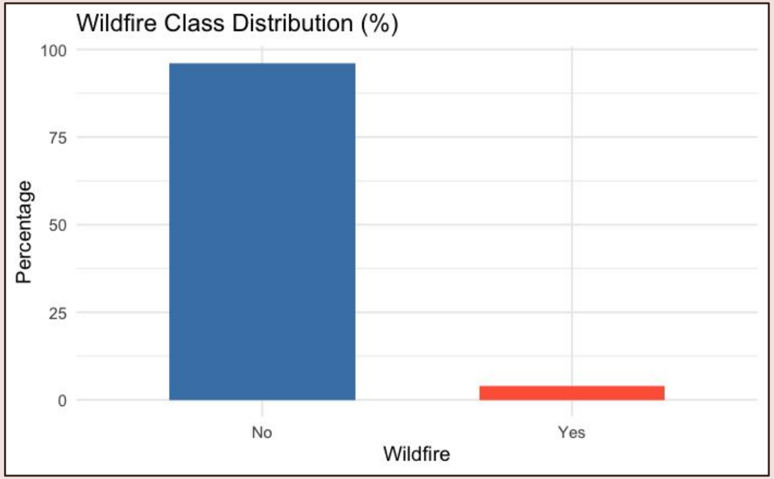
# Motivation

- Wildfires cause severe environmental and economic damage.
- Early detection helps in disaster prevention.
    - Climate factors influence wildfire behavior

# Objectives

- To predict wildfire occurrence (Yes/No)
- To use climate variables with KNN,random Forest,SVM,Decision Tree and Logistic Regression
- To evaluate each model performance on future data

# Data


- US Wildfire Dataset from Kaggle
- 2018–2019 data (56,025)
- Wildfire labels are highly imbalanced
- Makes harder for models to learn the wildfire “Yes” pattern without additional techniques such as SMOTE

# Predictive Variables

## Independent

latitude → spatial location

longitude → spatial location

pr → precipitation

rmax → maximum relative humidity

rmin → minimum relative humidity

sph → specific humidity

srad → solar radiation

tmmn → minimum temperature

tmmx → maximum temperature

vs → Wind speed

vpd → Vapor pressure deficit

fm100/fm1000 → Fuel moisture indices

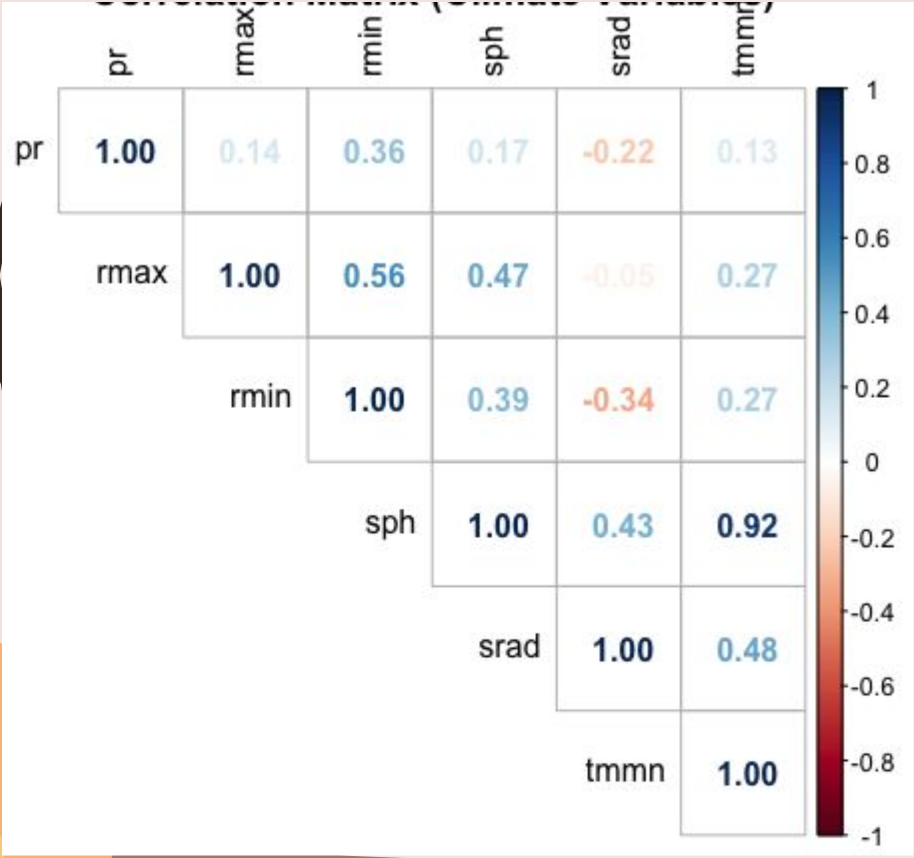
erc → Energy release component

vi → Burning index

etr, pet → Evapotranspiration

## Dependent

Wildfire (Yes / No)




- Most climate variables show low to moderate correlations with each other
- sph and tmmn show a strong positive correlation
- rmax–rmin and srad–tmmn also show moderate positive relationships
- Only a few weak negative correlations

Overall, predictors are not strongly collinear, making them suitable for multivariate modeling

# 02

# Modelling and Performances

# Machine Learning Models Applied

KNN

SVM

Random Forest

Decision Tree

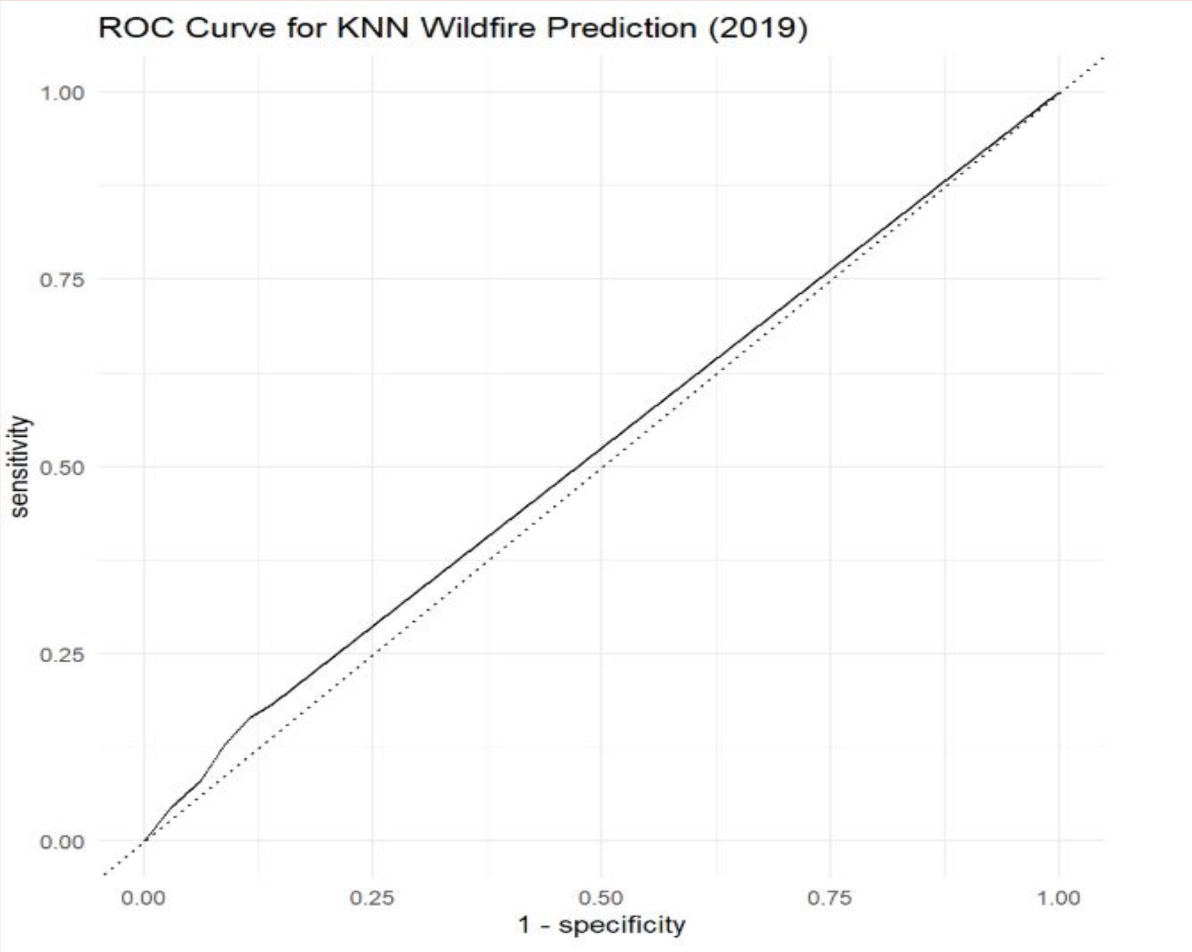
Logistic Regression

# KNN

- Binary classification Yes/No
- Training on 2018 data
- Testing on 2019 data
- Handling class imbalance with upsampling and SMOTE

# Evaluation Metrics

- Accuracy: 88%
- Recall (Wildfire): 12.8%
- F1-score: 7.6%
- Kappa: 0.024
- AUC: 0.523



ROC Curve for KNN Wildfire Prediction (2019)
sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - specificity

# KNN Interpretation

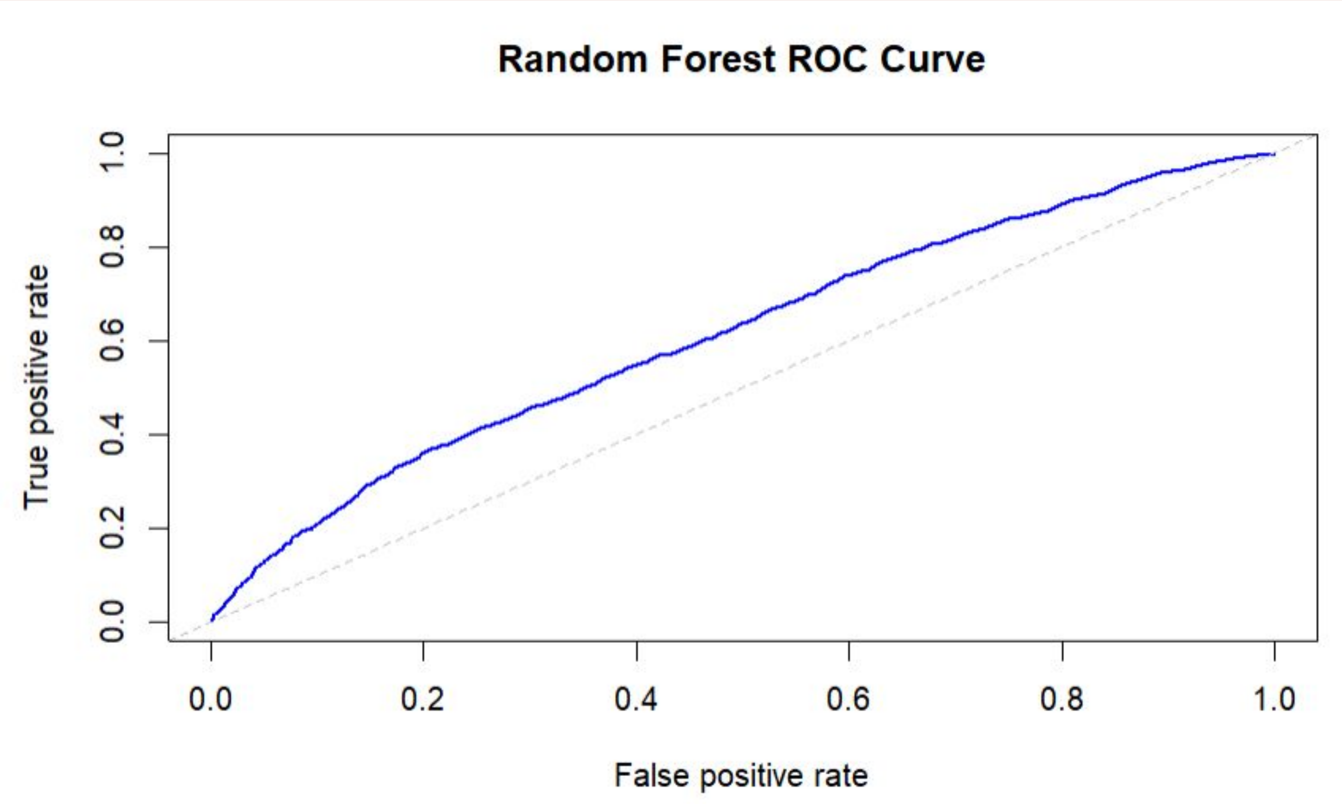
- KNN performs well on non-wildfire cases but poorly on wildfire detection
- Class imbalance strongly affects performance
- KNN is not reliable for wildfire prediction

# Random Forest

- Training on 2018 data
- Testing on 2019 data
- Handling class imbalance with upsampling and SMOTE

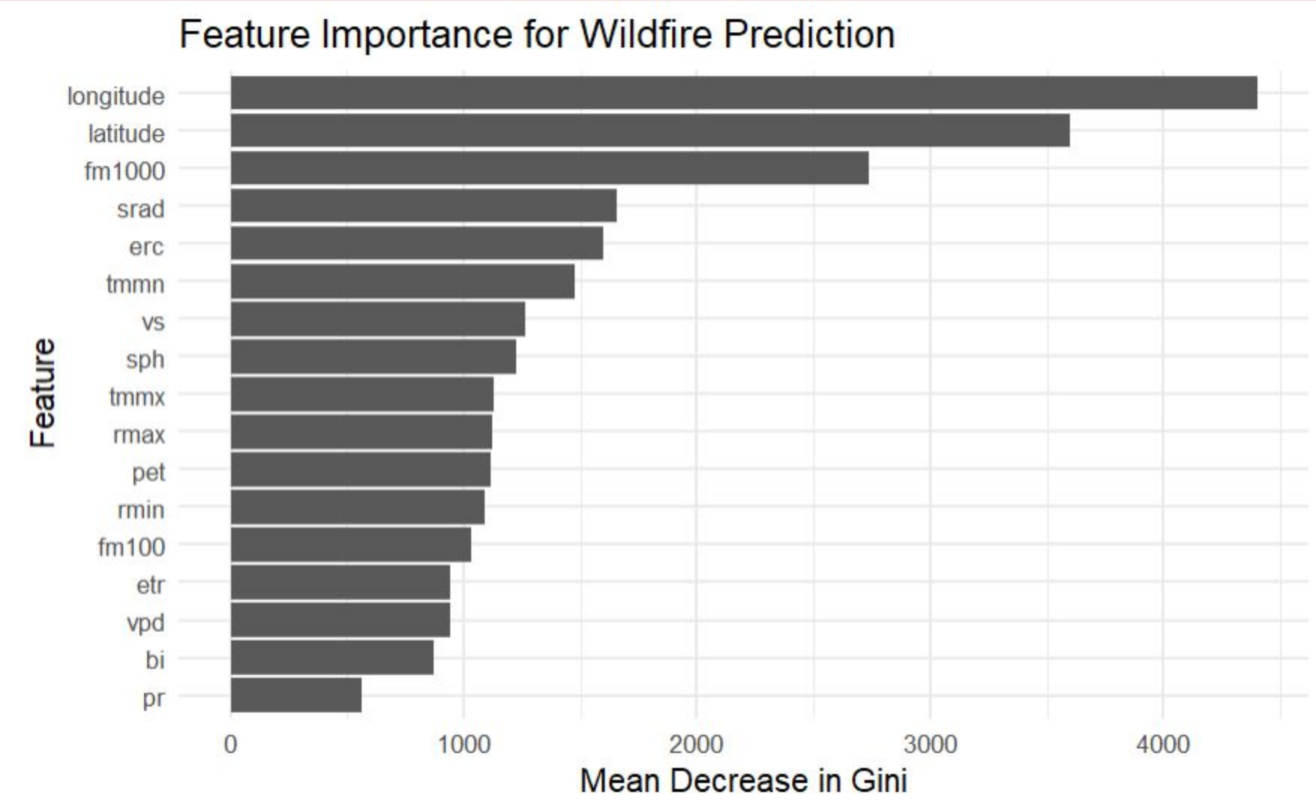
# Evaluation Metrics

- Accuracy: 93.47%
- Precision: 9.76%
- Recall (Wildfire): 8.62%
- F1-score: 9.15%
- AUC: 61.41%



Random Forest ROC Curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

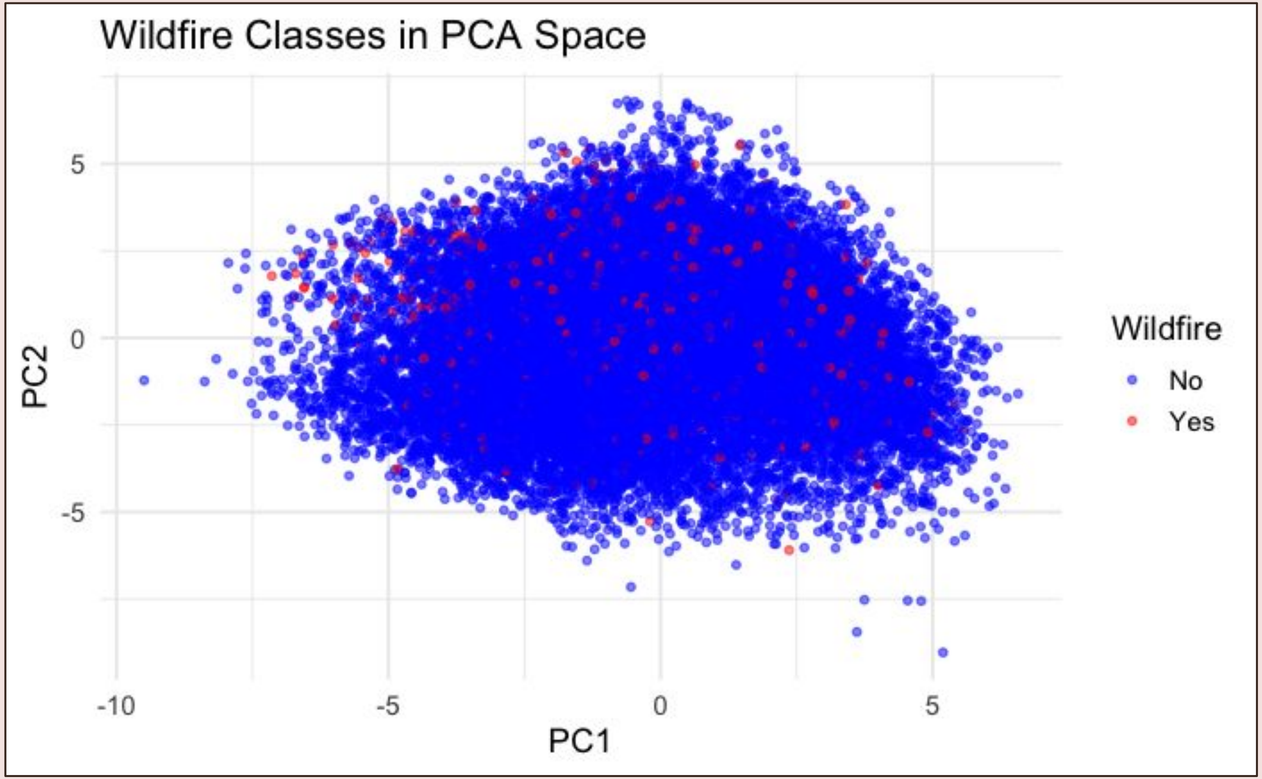
# Feature Importance



# Interpretation

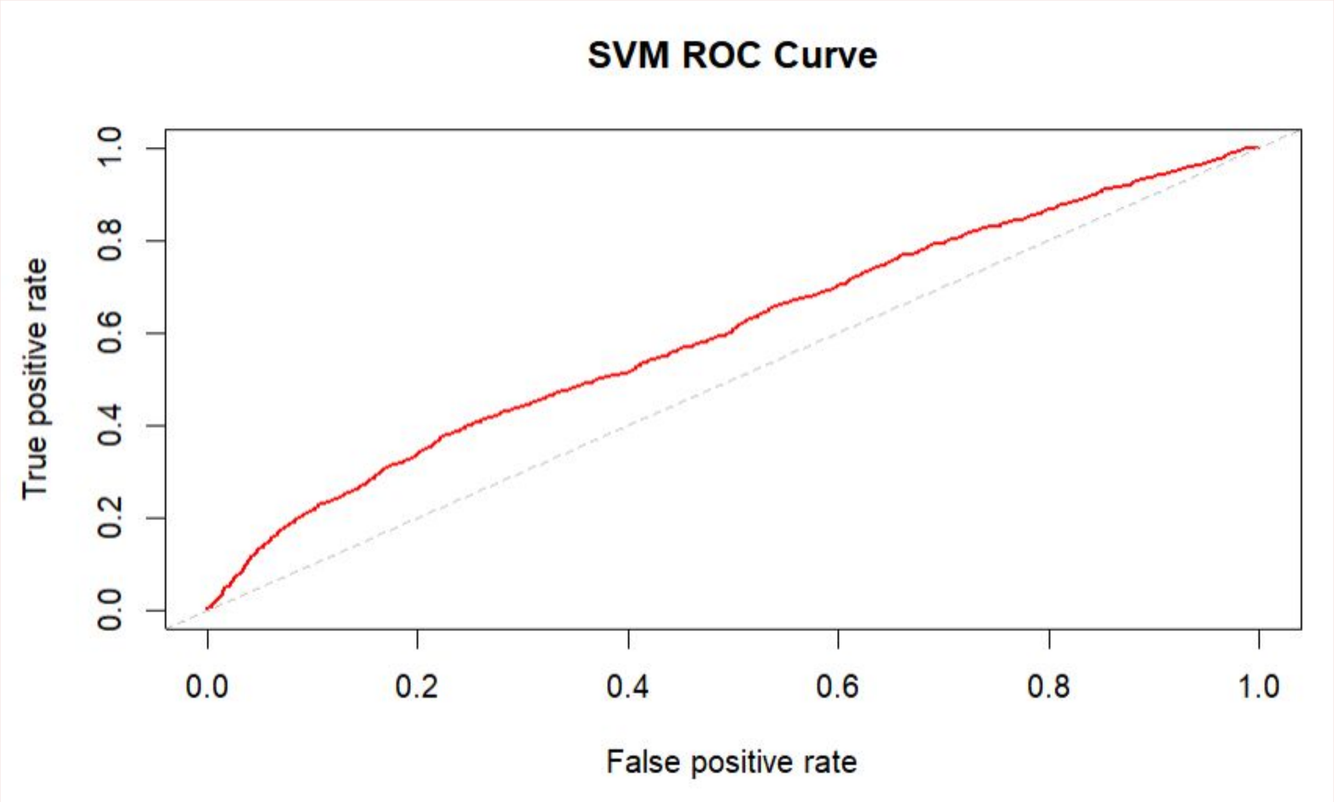
- Achieves high overall accuracy but fails to capture most wildfire cases.
- Generates many missed detections due to very low recall for the minority class.
- Limited ability to distinguish wildfire vs. non-wildfire situations, despite SMOTE.
- Not suitable for practical wildfire prediction where catching fires is critical.

# SVM


Wildfire Classes in PCA Space
5
0
-5
PC2
-10
-5
0
5
PC1
Wildfire
No
Yes

# Evaluation Metrics

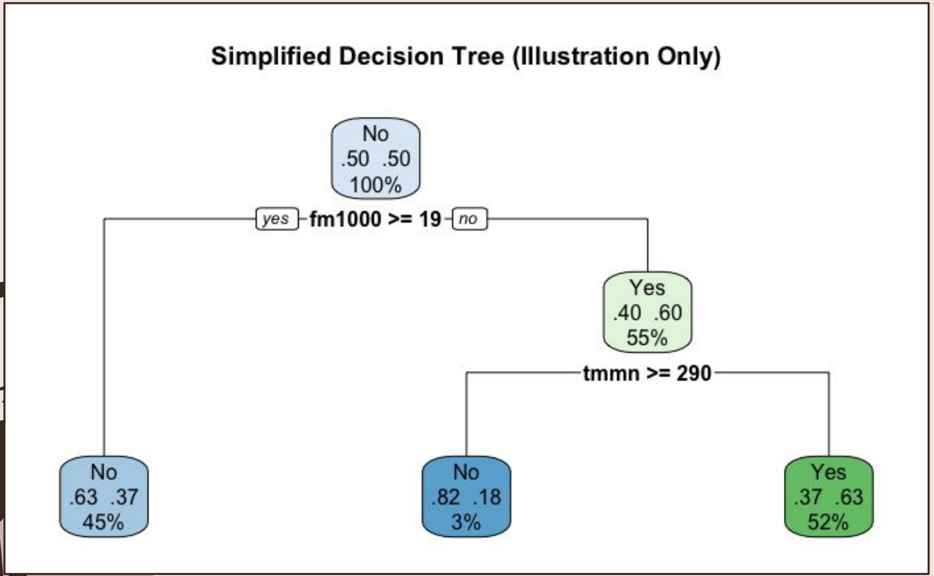
- Accuracy: 73.84%
- Precision: 6.00%
- Recall (Wildfire): 39.93%
- F1-score: 10.43%
- AUC: 59.61%


SVM ROC Curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

# Interpretation

- Shows moderate ability to distinguish wildfire vs. non-wildfire cases.
- Detects more wildfire events compared to other models, but with many false positives.
- Performance is limited by the strong class imbalance despite SMOTE.
- Useful for identifying potential wildfire risk, but not fully reliable on its own.

# Decision Tree




- Trained with 2018 data + SMOTE
- Pruned and depth-limited for interpretability
- Captures simple climate-based rules
- Recall improved but precision remained low
- Not as strong as Logistic/SVM but useful for understanding feature splits

# Logistic Regression

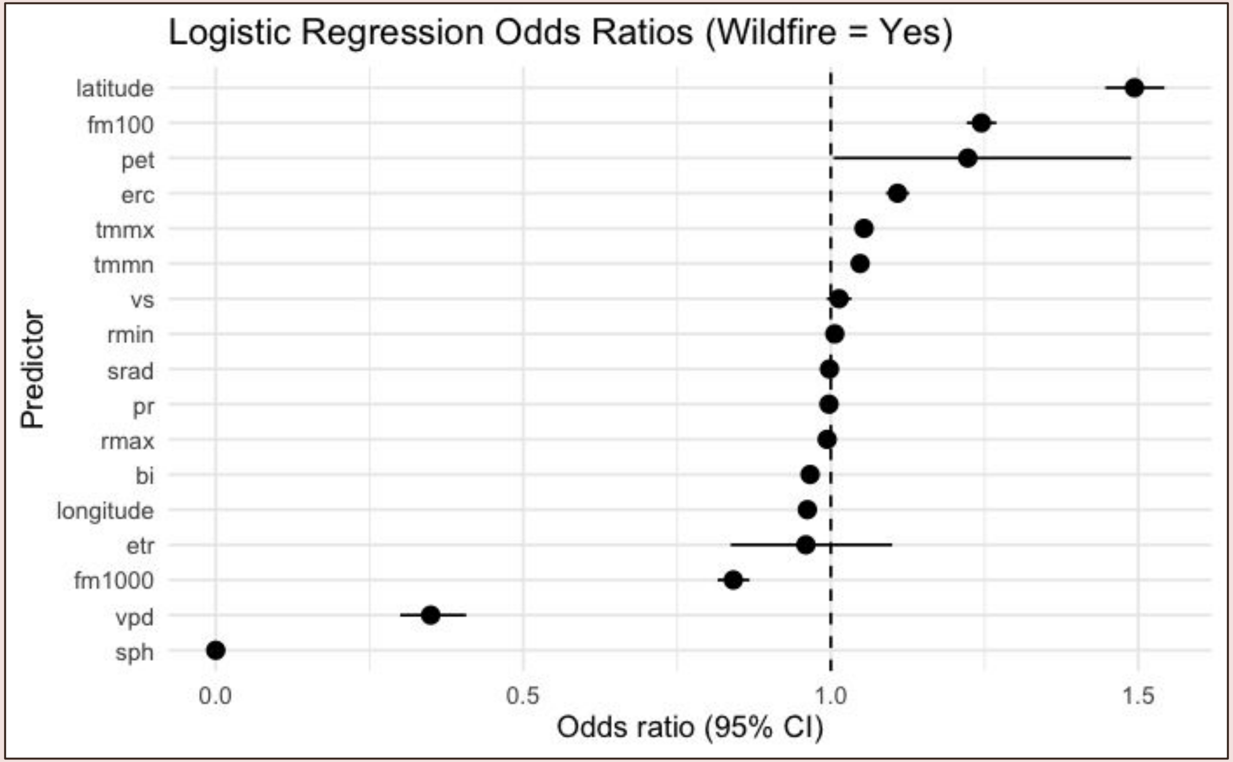
- Trained with 2018 data + SMOTE
- Evaluated at multiple cutoffs (0.1–0.5)
- Cutoff 0.5 provides the best balance (highest F1)
- Selected for producing stable recall without overfitting

Logistic Regression (SMOTE) performance across different cutoffs

| Cutoff | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.1 | 0.053 | 0.038 | 0.992 | 0.074 |
| 0.2 | 0.136 | 0.040 | 0.942 | 0.077 |
| 0.3 | 0.293 | 0.045 | 0.876 | 0.086 |
| 0.5 | 0.685 | 0.062 | 0.512 | 0.110 |

**Cutoff Decision:**

- 0.1 = high recall but extremely low F1
  → overfitting / too many false positives
- 0.5 = highest F1 + reasonable recall
  → best for reliable prediction


Logistic Regression Odds Ratios (Wildfire = Yes)
Predictor
latitude
fm100
pet
erc
tmmx
tmmn
vs
rmin
srad
pr
rmax
bi
longitude
etr
fm1000
vpd
sph
0.0
0.5
1.0
1.5
Odds ratio (95% CI)

04

# Conclusions

Best Model | Project Conclusion | Future Works

# Important Evaluation Metrics

### Most Important

## Recall

- Missing a wildfire is costly
- Measures how well the model captures rare "Yes" events

### Balance Check

## F1 Score

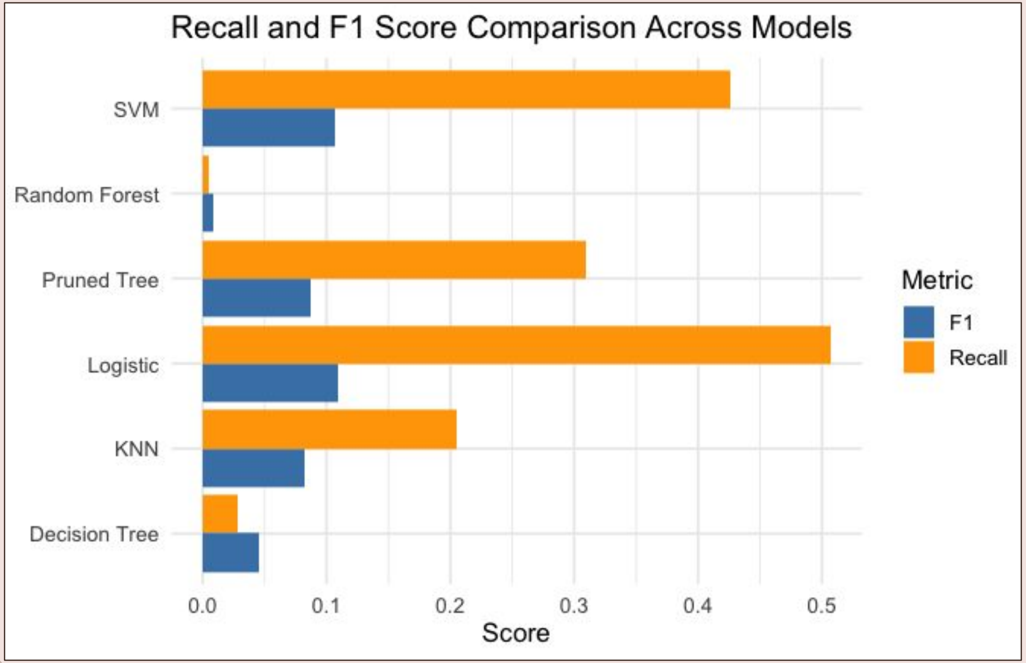
- High recall can increase false alarms
- F1 checks the balance between recall and precision

### Overall Model Quality

## AUC-ROC

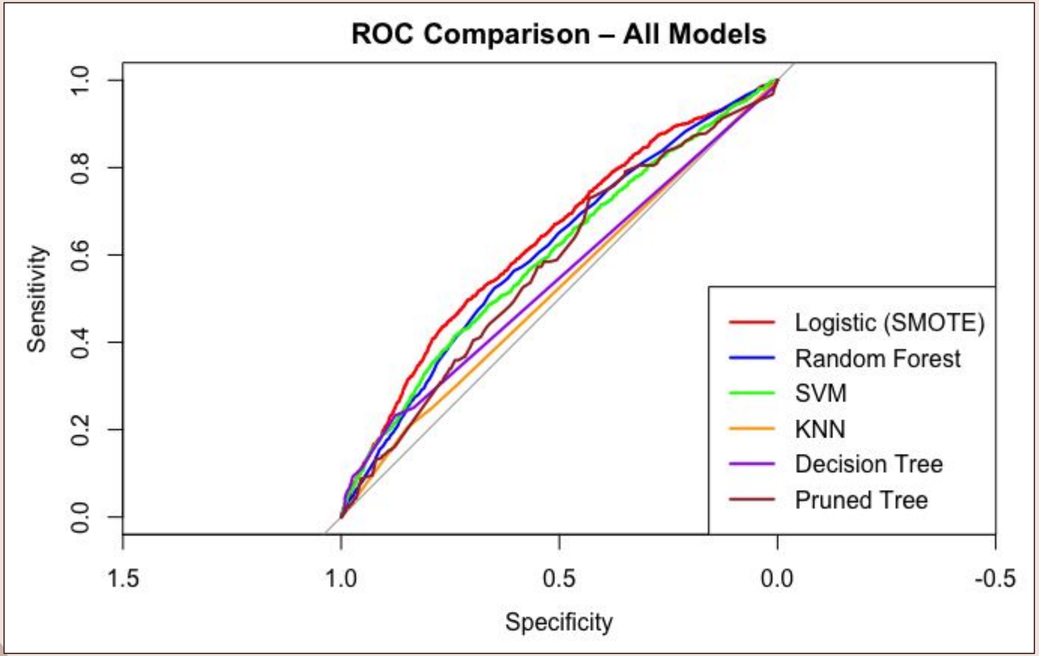
- Shows overall separation between wildfire and non-wildfire
- Less sensitive to thresholds and class imbalance

# Comparing Every Models




- *Logistic* and *SVM* show the highest recall among all models
- *Logistic* has the strongest recall, capturing the most wildfire events
- *F1 scores* for *Logistic* and *SVM* are nearly identical, but Logistic is slightly higher

# Comparing Every Models




- *SVM* and *Random Forest* show moderate AUC but *do not outperform Logistic*
- *Logistic* achieves the highest AUC, indicating the best overall discrimination ability

# Conclusion

- SMOTE improved all models by enabling minority-class learning
- Logistic Regression: the highest recall and best AUC, slightly higher F1 than SVM
- SVM performed similarly but was less stable across thresholds
- Tree-based models and KNN underperformed relative to Logistic Regression

Final Best Model: Logistic Regression (with SMOTE) for consistent wildfire detection

# Future Works

- **Add more features**: vegetation dryness, land cover, human activity
- Explore advanced **imbalance handling**
- **Test stronger models**: XGBoost, Gradient Boosting, ensembles
- Use spatial and temporal cross-validation for better generalization

# Thank You