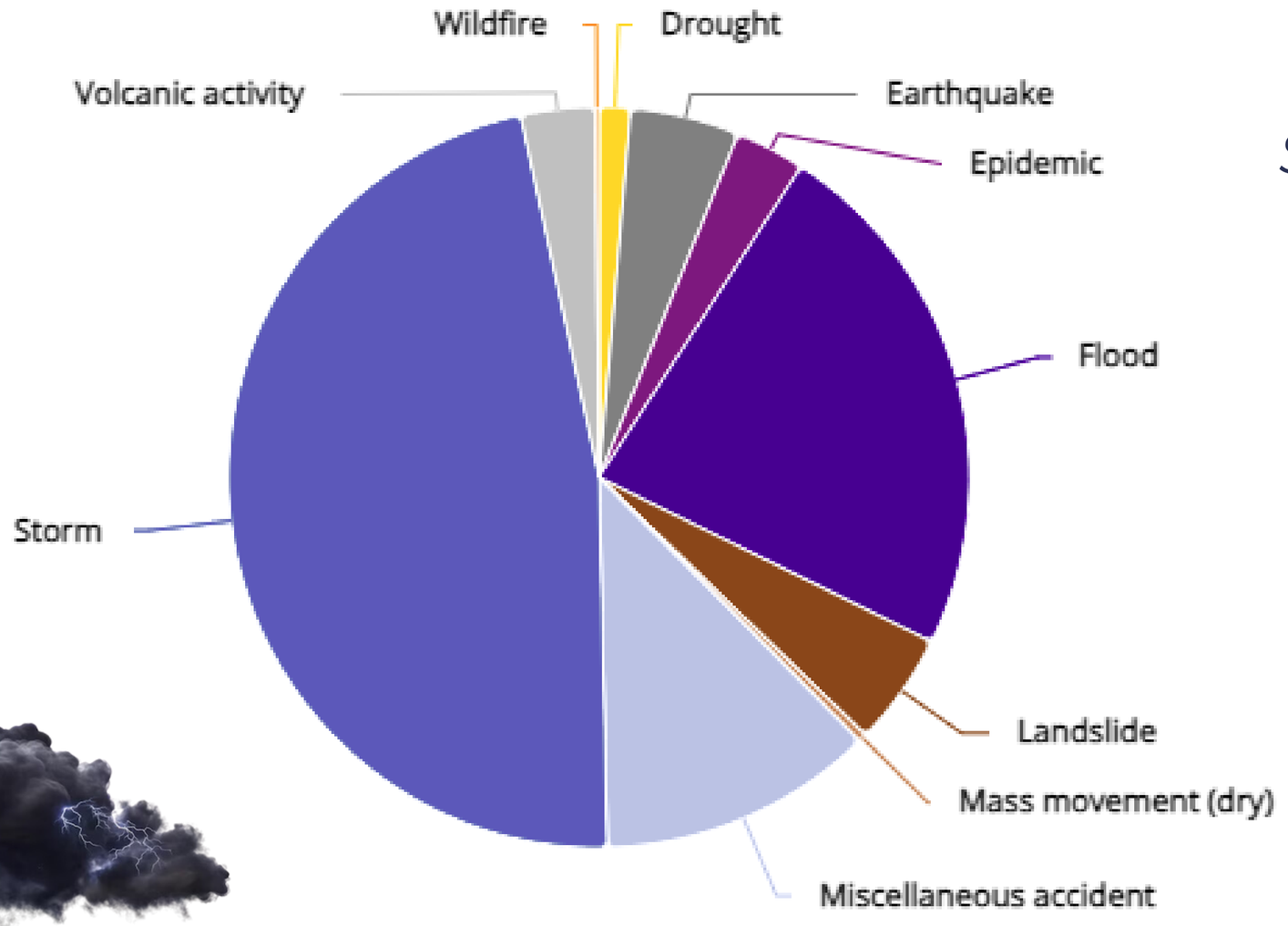
Average Annual Natural Hazard Occurrence for 1980-2020



Source: https://climateknowledgeportal.worldbank.org/country/philippines/vulnerability

Since 1990, the Philippines has faced 565 natural disasters

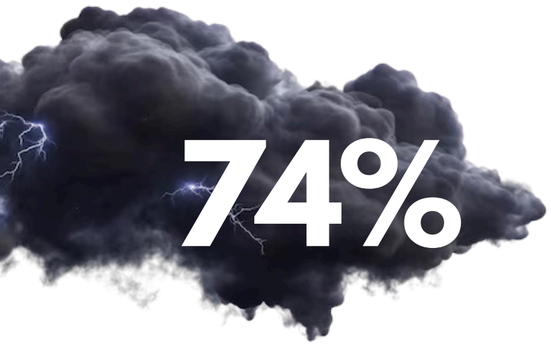
70,000 deaths

$23 Billion in damages

Motivation

Convective Storms

- Convective storms cause significant risks to public safety, infrastructure, and economic stability.
- Vulnerability amplified by **urban heat island effects, monsoons,** and **rapid urbanization.**



74%

*of Philippine's population is exposed to natural disaster hazards including **storms**.*

ThorM:

A Machine Learning-Based Classification of Convective **Storm Risk Levels** using Climatological Data in Metro Manila, Philippines

Predicting Storm Risk Levels

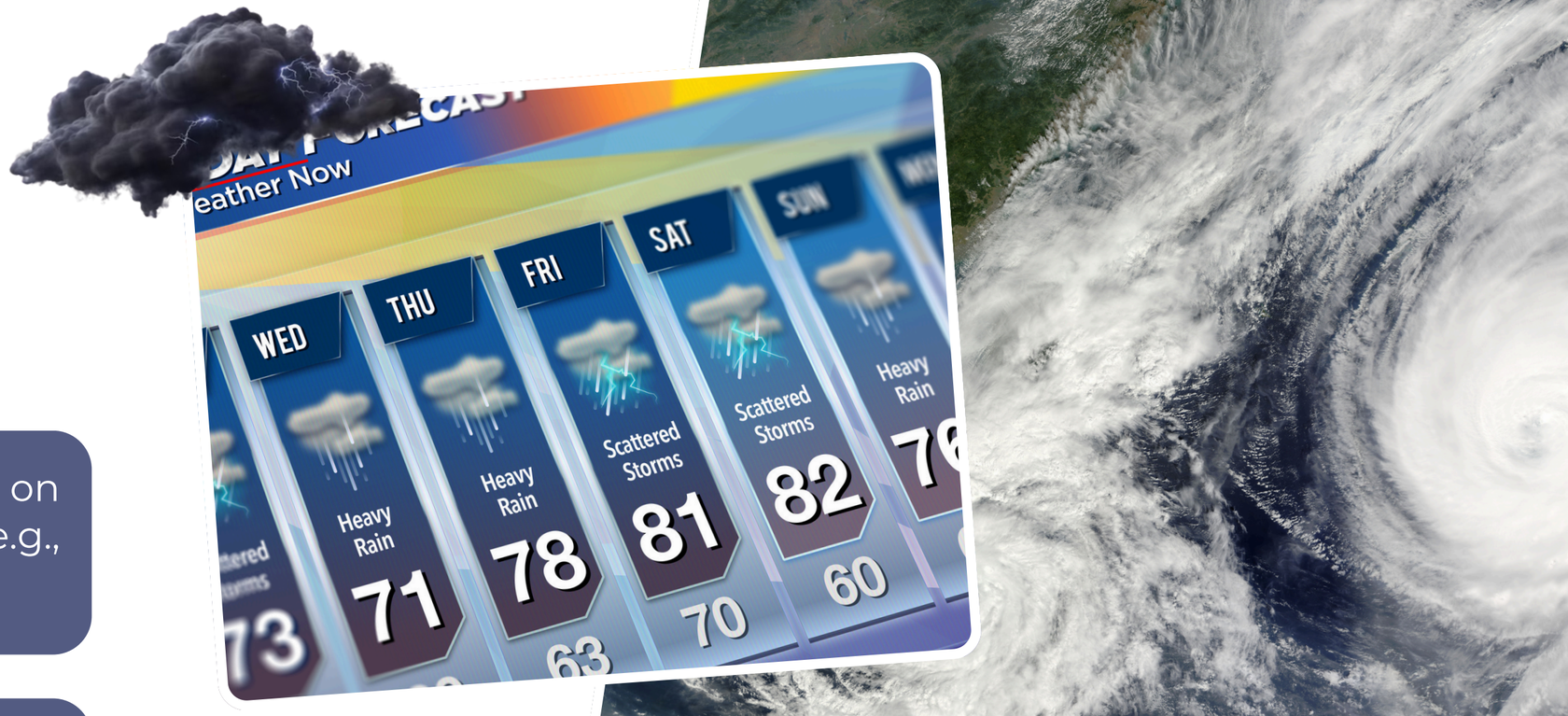
CELIS, PORRAS, SALINAS

Challenges

in Forecasting

Traditional methods rely on computationally heavy indices (e.g., CAPE, CIN).

Lack of localized adaptability for urban areas.



Objectives

Generally, our project aims to develop a machine learning-based framework for classifying convective storm risk levels in Metro Manila, Philippines.

Specific Objectives

- *Apply and compare machine learning models (SVM, KNN, Decision Trees, Random Forests, XGBoost) for storm risk classification.*
- *Identify key weather variables influencing storm risk levels using Recursive Feature Elimination (RFE).*
- *Optimize model performance using hyperparameter tuning techniques.*

Methodology

Dataset Overview

Source: Visual Crossing Corporation

Coverage: Quezon City, Metro Manila (Oct 1, 2020 - Oct 1, 2024)

Total Entries: 1,462
Features: 33

Target Variable

Name: ***severerisk***

Type: Float64

Non-null entries: 996

Element	Description
name	Location name
datetime	Date and time of observation
tempmax	Maximum temperature recorded
tempmin	Minimum temperature recorded
temp	Average temperature
feelslikemax	Maximum "feels like" temperature
feelslikemin	Minimum "feels like" temperature
feelslike	Average "feels like" temperature
dew	Dew point
humidity	Humidity percentage
precip	Precipitation amount
precipprob	Probability of precipitation
precipcover	Coverage of precipitation
preciptype	Type of precipitation (e.g., rain, snow)
snow	Snow amount
snowdepth	Snow depth
windgust	Maximum wind gust speed
windspeed	Average wind speed
winddir	Wind direction
sealevelpressure	Sea level pressure
cloudcover	Cloud cover percentage
visibility	Visibility distance
solarradiation	Solar radiation level
solarenergy	Solar energy level
uvindex	UV index
severerisk	Severe weather risk level
sunrise	Time of sunrise
sunset	Time of sunset
moonphase	Moon phase percentage
conditions	General weather conditions (e.g., cloudy, clear)
description	Detailed weather description
icon	Weather icon representation
stations	Station data source(s)

Dataset Overview

- **Source:** Visual Crossing Corporation
- **Coverage:** Quezon City, Metro Manila (Oct 1, 2020 - Oct 1, 2024)
- **Total Entries:** 1,462
  Features: 33

Target Variable

Name: ***severerisk***

Type: Float64

Non-null entries: 996

snowdepth	Snow depth
windgust	Maximum wind gust spe
windspeed	Average wind speed
winddir	Wind direction
sealevelpressure	Sea level pressure
cloudcover	Cloud cover percentage
visibility	Visibility distance
solarradiation	Solar radiation level
solarenergy	Solar energy level
uvindex	UV index
severerisk	Severe weather risk level
sunrise	Time of sunrise
sunset	Time of sunset
moonphase	Moon phase percentage
conditions	General weather conditio
description	Detailed weather descript

Exploratory Data Analysis

Target Variable

severerisk

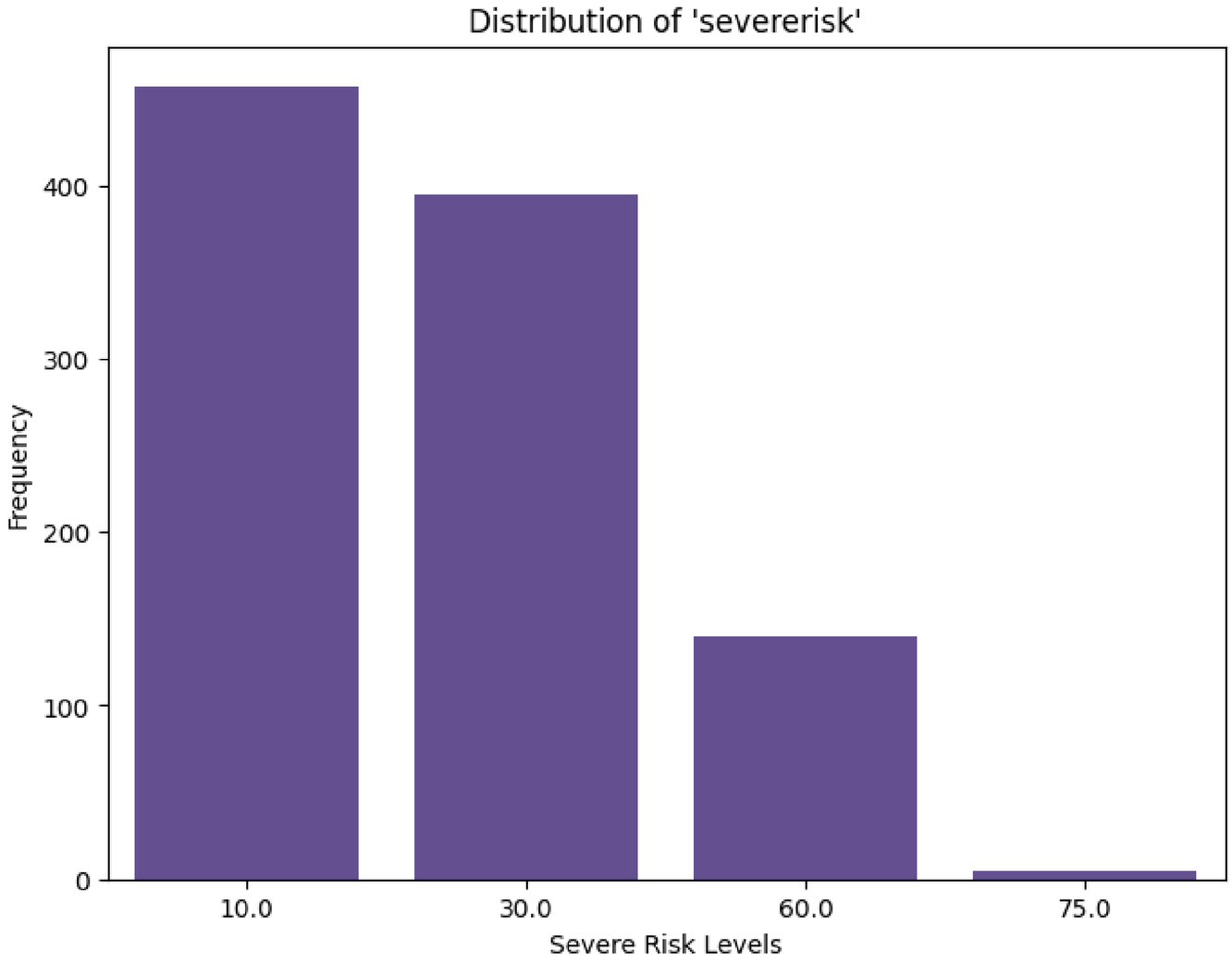
- Continuous variable (0-100)
- Derived from CAPE, CIN, predicted precipitation, and wind

Imbalance:

- Most values clustered around 10-30
- Rare extreme values (60-70)

Categories:

- **Low:** < 30
- **Moderate:** 30-69
- **High:** ≥ 70



Exploratory Data Analysis

Correlations

Highly Correlated Features

- *dew:* 0.73
- month_sin: 0.58
- feelslike: 0.55
- daylight_hours: 0.54

Distribution

Heavily Skewed Features
(Skewness > |1|)

- *precip*: 8.005
- *windgust*: 6.793
- *windspeed*: 4.112

Outliers

- Significant outliers in **tempmin, visibility, windgust,** and **sealevelpressure.**
- Meaningful outliers retained to reflect severe weather conditions.

Steps in

Preprocessing

Handling Missing Values | ***Dropping Irrelevant Features*** | ***Feature Engineering & Encoding*** | ***Handling Skewness*** | ***Data Preparation*** | ***SMOTE*** | ***Feature Scaling***

- Target Variable **(severerisk)**: Rows with missing values dropped.
- **preciptype**: Missing values replaced with "None," encoded as binary (Rain=1, None=0).

Steps in

Preprocessing

Handling Missing Values → ***Dropping Irrelevant Features*** → ***Feature Engineering & Encoding*** → ***Handling Skewness*** → ***Data Preparation*** → ***SMOTE*** → ***Feature Scaling***

- Removed: name, stations, description, icon, snow, snowdepth.
- Retained only one among **highly correlated** features to reduce redundancy (e.g., tempmin over tempmax, feelslike over feelslikemin).

Steps in

Preprocessing

Handling Missing Values — ***Dropping Irrelevant Features*** — ***Feature Engineering & Encoding*** — ***Handling Skewness*** — ***Data Preparation*** — ***SMOTE*** — ***Feature Scaling***

- Derived cyclical features **(month_sin, month_cos)** to capture seasonal trends.
- Added **daylight_hours** (difference between sunrise and sunset).
- Target variable **(severerisk_encoded)** categorized into Low, Moderate, and High risk levels, numerically encoded (0, 1, 2).
- Categorical feature **conditions** one-hot encoded.

Steps in

Preprocessing

Handling Missing Values — ***Dropping Irrelevant Features*** — ***Feature Engineering & Encoding*** — ***Handling Skewness*** — ***Data Preparation*** — ***SMOTE*** — ***Feature Scaling***

- Skewed features (precip, windgust, windspeed) normalized using **Yeo-Johnson** transformation to improve distribution and reduce bias.

Steps in

Preprocessing

Handling Missing Values → ***Dropping Irrelevant Features*** → ***Feature Engineering & Encoding*** → ***Handling Skewness*** → ***Data Preparation*** → ***SMOTE*** → ***Feature Scaling***

- Stratified split: **80% training, 20% testing**, maintaining class distribution.

Steps in

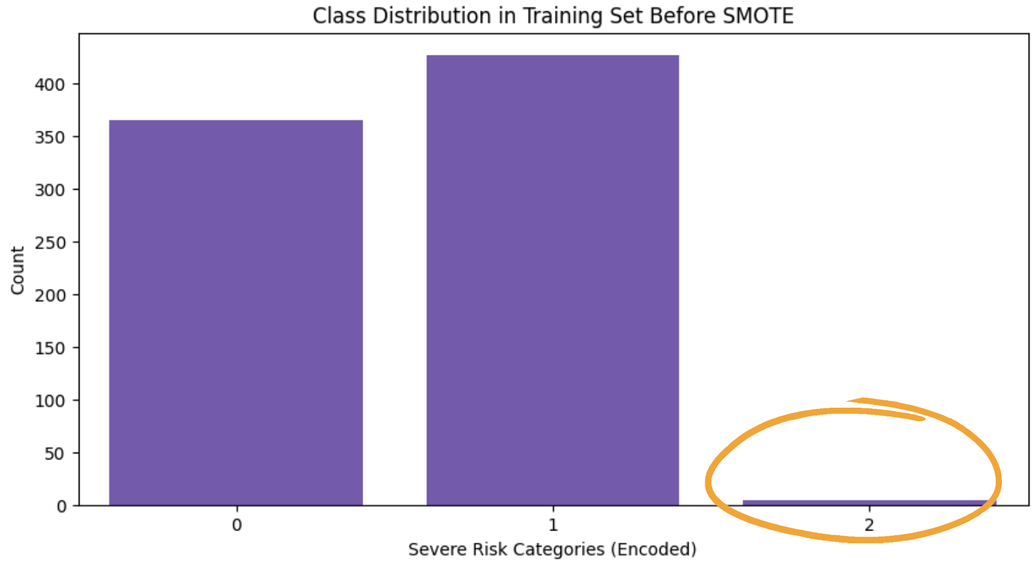
Preprocessing

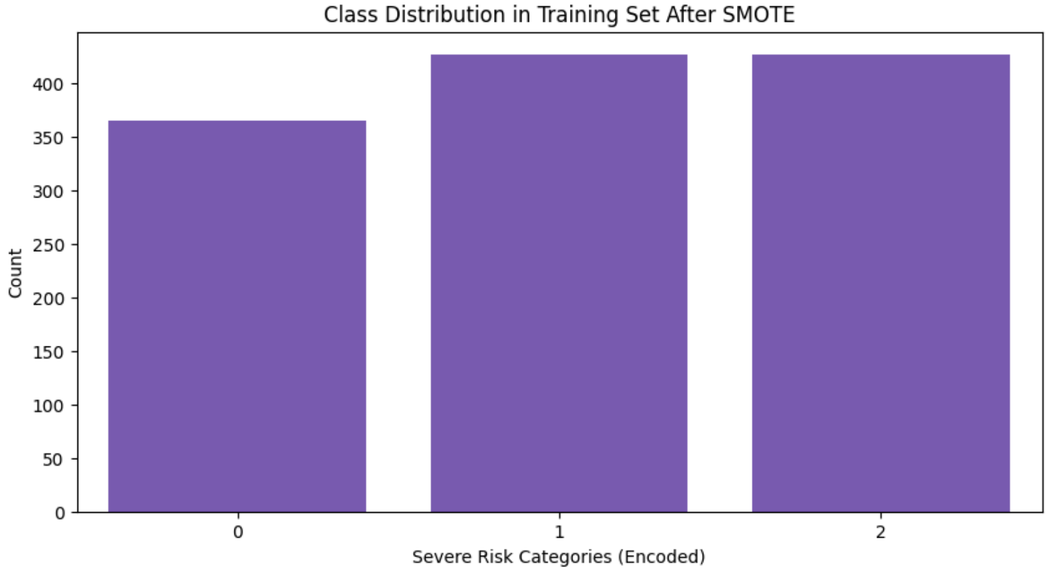
- Applied SMOTE on the training set to **generate synthetic samples** for the underrepresented High-risk class, avoiding overfitting.

Steps in

Preprocessing

- *Handling Missing Values*
- *Dropping Irrelevant Features*
- *Feature Engineering & Encoding*
- *Handling Skewness*
- *Data Preparation*
- *SMOTE*
- *Feature Scaling*





Steps in

Preprocessing

Handling Missing Values → ***Dropping Irrelevant Features*** → ***Feature Engineering & Encoding*** → ***Handling Skewness*** → ***Data Preparation*** → ***SMOTE*** → ***Feature Scaling***

- **Standardized numerical features** using StandardScaler to ensure compatibility with ML algorithms, particularly SVM.

Classification

Model Training and Testing

Performance

Metrics

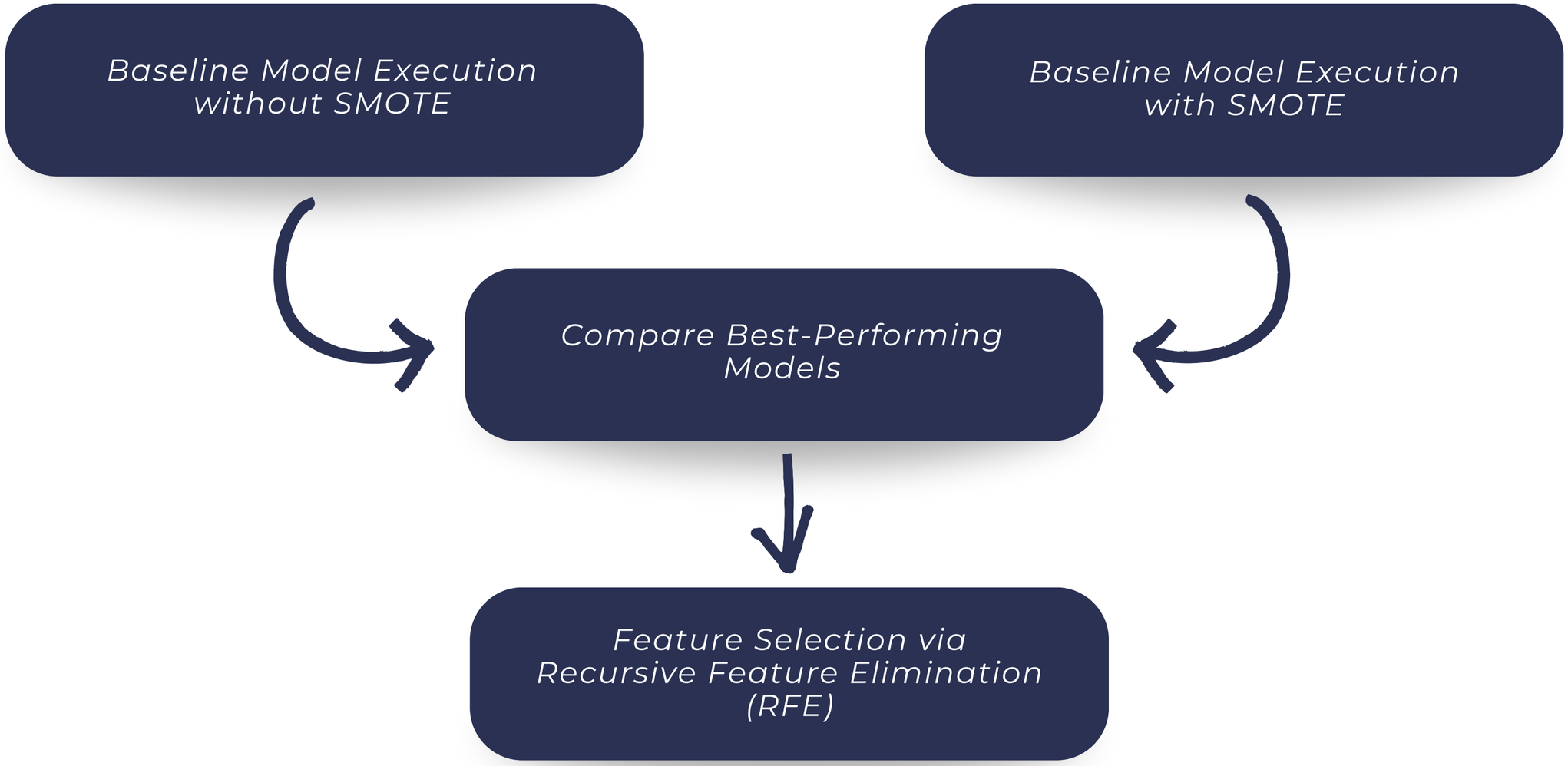
Accuracy

Precision

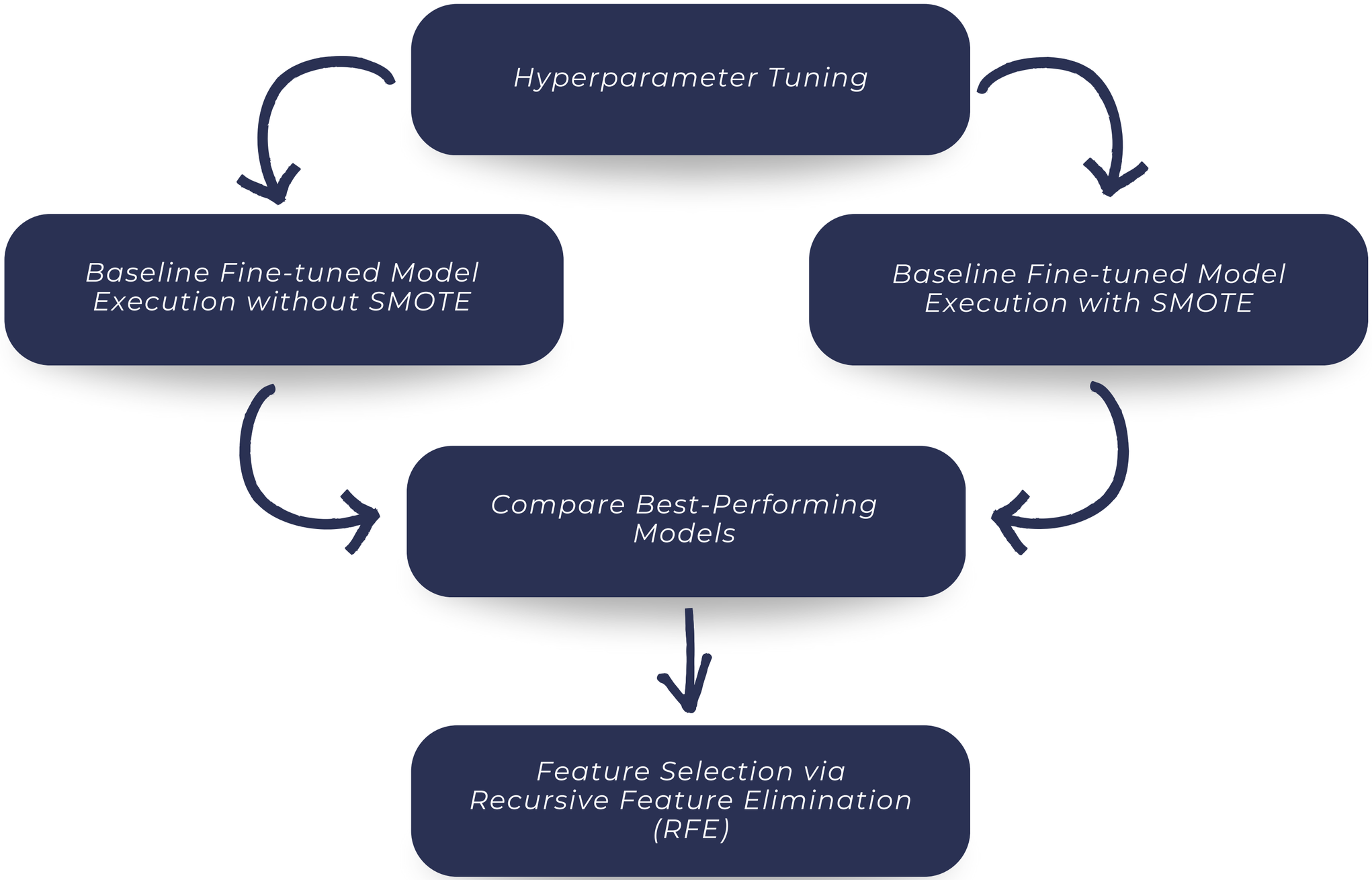
Recall

F1-Score

Confusion Matrix


Baseline Model Execution without SMOTE
Baseline Model Execution with SMOTE
Compare Best-Performing Models
Feature Selection via Recursive Feature Elimination (RFE)

Hyperparameter Tuning


Hyperparameter Tuning
Baseline Fine-tuned Model Execution without SMOTE
Baseline Fine-tuned Model Execution with SMOTE
Compare Best-Performing Models
Feature Selection via Recursive Feature Elimination (RFE)

Results and Discussion

Baseline Model Execution

Performance Results

Baseline Model Execution without SMOTE

Baseline Model Execution with SMOTE

		ML Models				
	Metric	**RF**	**DT**	**SVM**	**KNN**	**XGBoost**
without SMOTE	Accuracy	0.8350	0.7700	0.8300	0.8300	**0.8600**
	Precision	0.8348	0.7782	0.8366	0.8306	**0.8623**
	Recall	0.8350	0.7700	0.8300	0.8300	**0.8600**
	F1-Score	0.8330	0.7675	0.8276	0.8280	**0.8578**
with SMOTE	Accuracy	**0.8800**	0.8350	0.8550	0.8000	0.8600
	Precision	**0.8773**	0.8311	0.8603	0.8255	0.8571
	Recall	**0.8800**	0.8350	0.8550	0.8000	0.8600
	F1-Score	**0.8774**	0.8326	0.8570	0.8123	0.8574

Random Forest

with Selected Features

Feature Selection via Recursive Feature Elimination (RFE)

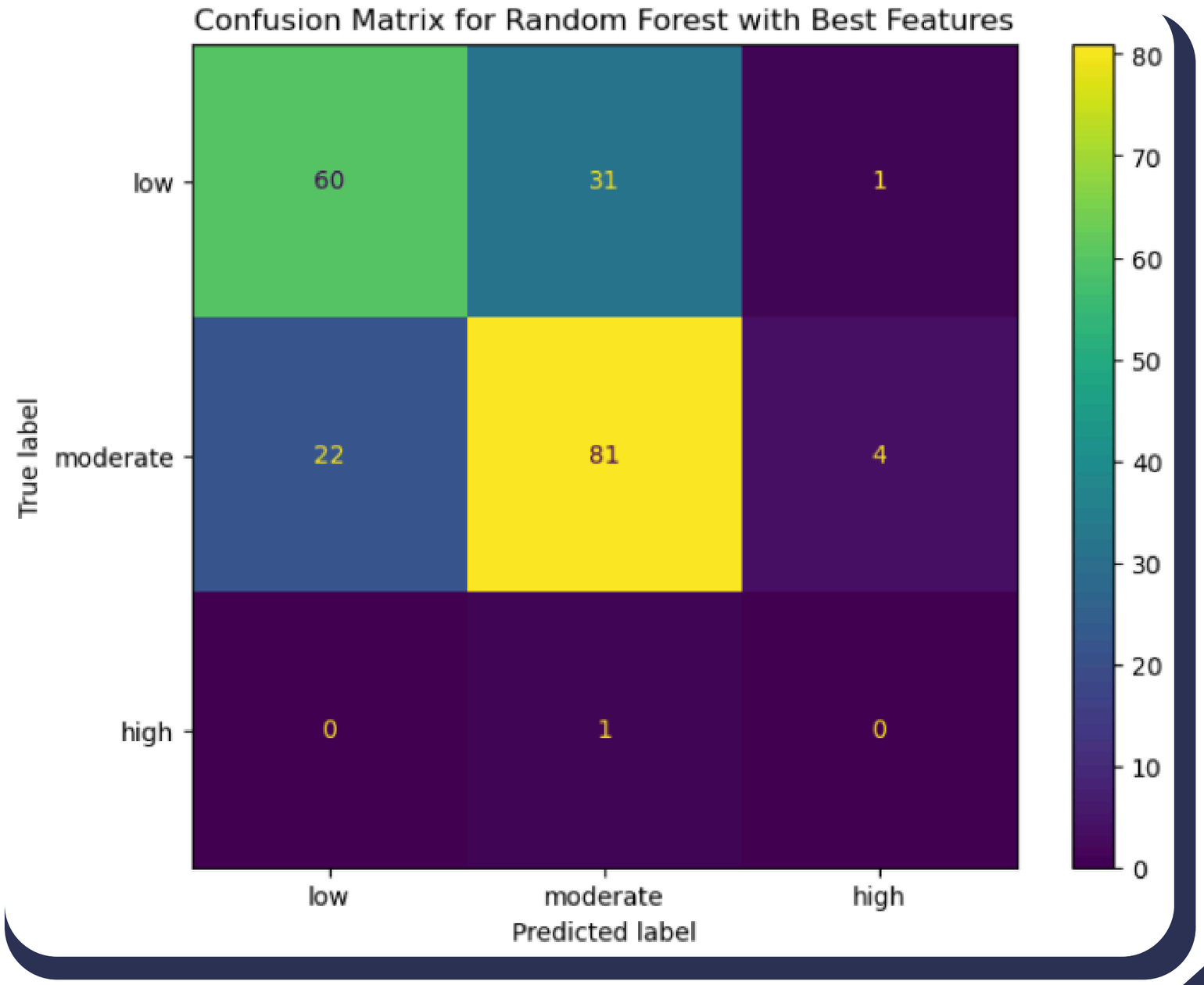
10 Selected Features

88.50% Accuracy

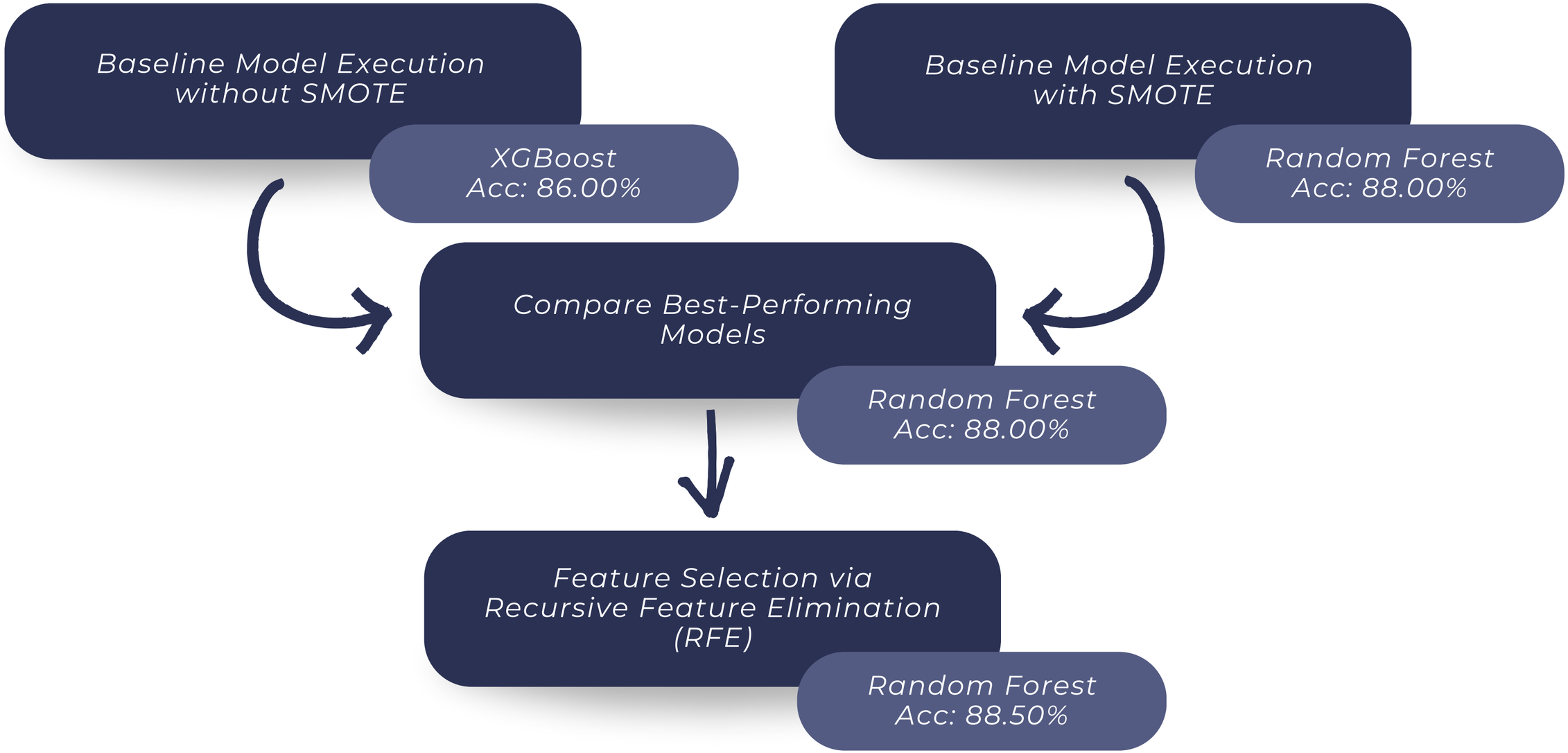
10 Selected Features

Feature Selection and Importance via Recursive Feature Elimination (RFE)

	Random Forest	
1	**dew**	**0.21959412**
2	**feelslike**	**0.186166906**
3	**daylight_hours**	**0.110892954**
4	month_sin	0.099351621
5	humidity	0.080786904
6	precipcover	0.072265163
7	tempmin	0.065567475
8	precip	0.064719914
9	month_cos	0.055537098
10	sealevelpressure	0.045117845



88.50%
Accuracy


Baseline Model Execution without SMOTE
XGBoost Acc: 86.00%
Baseline Model Execution with SMOTE
Random Forest Acc: 88.00%
Compare Best-Performing Models
Random Forest Acc: 88.00%
Feature Selection via Recursive Feature Elimination (RFE)
Random Forest Acc: 88.50%

Hyperparameter Tuning

Hyperparameter Tuning

Cross-Validated Accuracy is based on the Training Set

Model	Best Hyperparameters	Cross-Validated Accuracy
Random Forest	max_depth=None, min_samples_leaf=4, min_samples_split=2, n_estimators=100	93.68%
Decision Tree	max_depth=10, min_samples_leaf=1, min_samples_split=10	90.57%
SVM	max_depth=10, min_samples_leaf=1, min_samples_split=10	92.78%
KNN	metric=minkowski, n_neighbors=5, weights=distance	90.32%
XGBoost	colsample_bytree=0.8, learning_rate=0.1, max_depth=5, n_estimators=100, subsample=0.8	93.93%

Baseline Fine-Tuned Model Execution

Performance Results

Baseline Model Execution without SMOTE

Baseline Model Execution with SMOTE

		ML Models				
	Metric	**RF**	**DT**	**SVM**	**KNN**	**XGBoost**
without SMOTE	Accuracy	0.8300	0.7800	0.8300	0.8300	**0.8450**
	Precision	0.8291	0.7861	0.8343	0.8306	**0.8448**
	Recall	0.8300	0.7800	0.8300	0.8300	**0.8450**
	F1-Score	0.8280	0.7777	0.8278	0.8280	**0.8430**
with SMOTE	Accuracy	0.8650	0.8350	0.8500	0.8050	**0.8750**
	Precision	0.8627	0.8362	0.8605	0.8310	**0.8741**
	Recall	0.8650	0.8350	0.8500	0.8050	**0.8750**
	F1-Score	0.8623	0.8345	0.8541	0.8174	**0.8721**

Fine-tuned XGBoost

with Selected Features

Feature Selection via Recursive Feature Elimination (RFE)

19 Selected Features

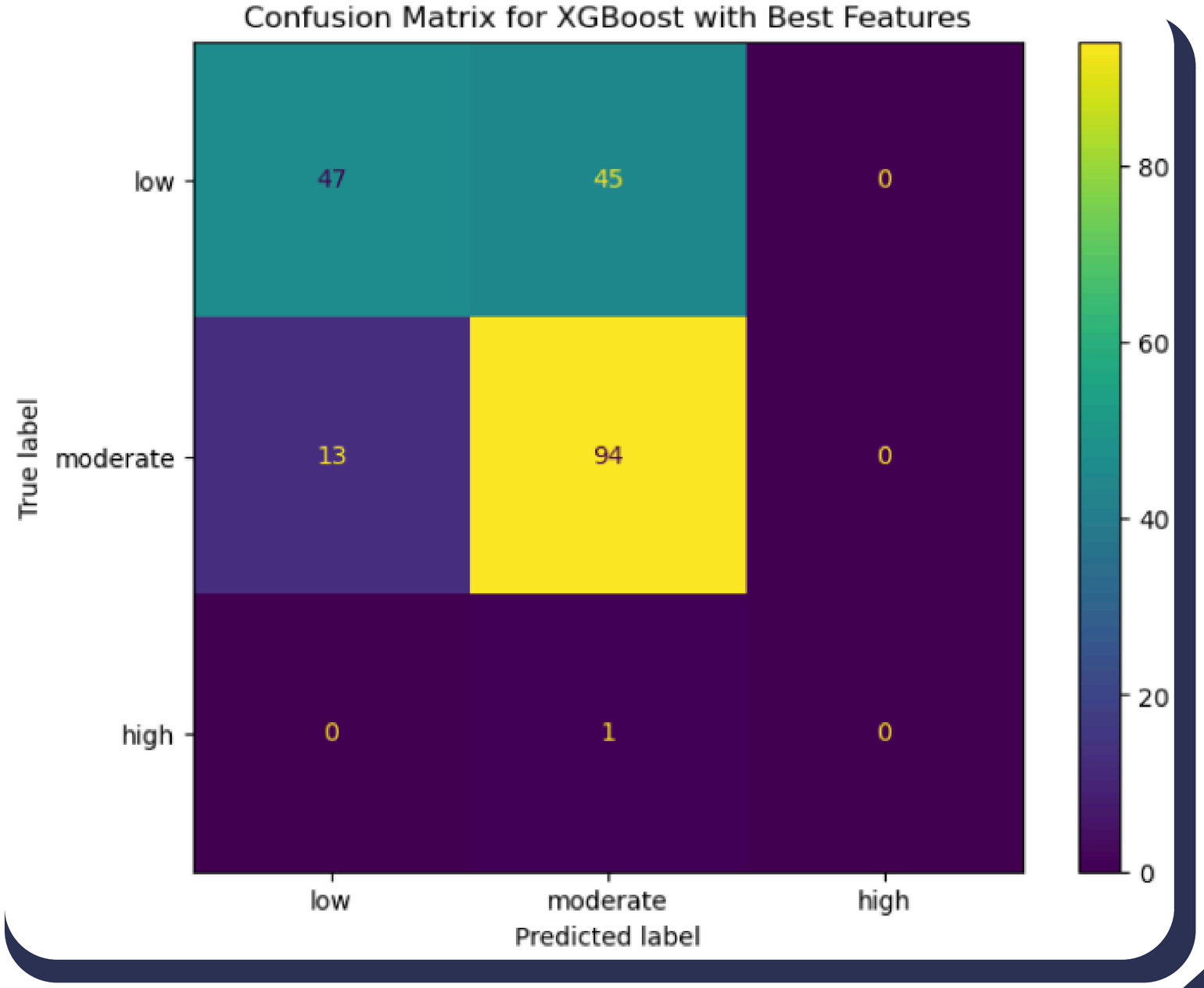
89.00% Accuracy

19 Selected Features

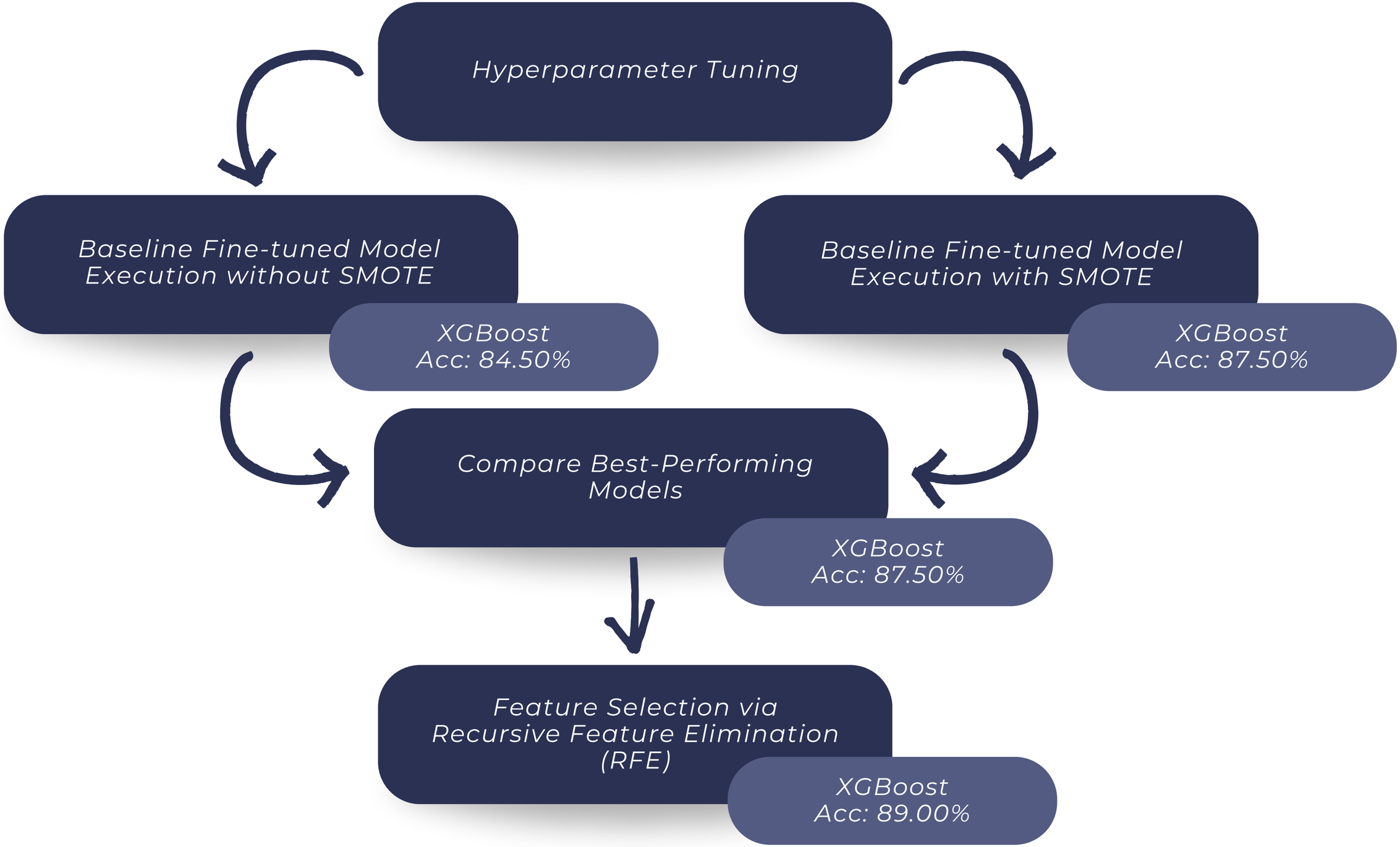
Feature Selection via Recursive Feature Elimination (RFE)

XGBoost		
1	**feelslike**	**0.13852261**
2	**precipprob**	**0.12824851**
3	**daylight_hours**	**0.12620859**
4	dew	0.1133081
5	month_sin	0.09604357
6	precipcover	0.082659155
7	precip	0.05299959
8	month_cos	0.049710866
9	humidity	0.047616098
10	tempmin	0.023141792

XGBoost (cont.)		
11	cloudcover	0.020571383
12	windgust	0.018897805
13	sealevelpressure	0.018142272
14	windspeed	0.016690673
15	solarenergy	0.015433878
16	solarradiation	0.014509254
17	winddir	0.013792287
18	visibility	0.013230527
19	uvindex	0.010273061



89.00%
Accuracy


Hyperparameter Tuning
Baseline Fine-tuned Model Execution without SMOTE
XGBoost
Acc: 84.50%
Baseline Fine-tuned Model Execution with SMOTE
XGBoost
Acc: 87.50%
Compare Best-Performing Models
XGBoost
Acc: 87.50%
Feature Selection via Recursive Feature Elimination (RFE)
XGBoost
Acc: 89.00%

Conclusion

Key Findings

- **Random Forest** performed well with a baseline accuracy of **88.50%** after RFE.
- **XGBoost** performed well with a hyperparameter-tuned accuracy of **89.0%** after RFE.
- In EDA, highly correlated features with storm risk include **dew**, **month_sin**, **feelslike**, **daylight_hours**
- After RFE of Random Forest, identified critical features influencing storm risk: **dew**, **feelslike**, **daylight_hours**
- After RFE of XGBoost, identified critical features influencing storm risk: **feelslike, precibprob, daylight_hours**

Conclusion

Implications

- Even without CAPE and CIN, the models still achieve high accuracy (88.5%-89.0%), demonstrating that these simpler variables are **sufficient proxies** for capturing storm risk dynamics.
- Key features like **dew, feelslike,** and **daylight_hours** consistently showed strong correlations with storm risk, indicating their significance in predictive modeling.
- By using commonly available data, storm risk prediction becomes **accessible** to regions with limited resources or less sophisticated weather monitoring infrastructure.

Conclusion

Applications

- **Enables** enhanced **storm preparedness** through predictive insights.
- **Supports** early warning systems and resource allocation for disaster management.
- Adaptable for urban planning and scalable for **real-time storm risk monitoring.**
- **Computationally efficient**, with potential for use in broader urban settings.

Conclusion

Future Work

- **Expand dataset coverage** across regions and timeframes.
- **Explore** ensemble learning and deep learning for higher predictive accuracy.
- **Integrate** satellite and socio-economic data for a more holistic risk assessment.
- **Test** real-world deployment with real-time system updates.

Thank You