

AI for Social Science Research

Ali Ünlü, PhD

Research Scientist

School of Education and Human Development

University of Virginia

aliunlu@virginia.edu

Session Organization



1st session: AI for Social Science Research



2nd session: Social Media Research



Materials:

Slides

Codes for feature selection methods

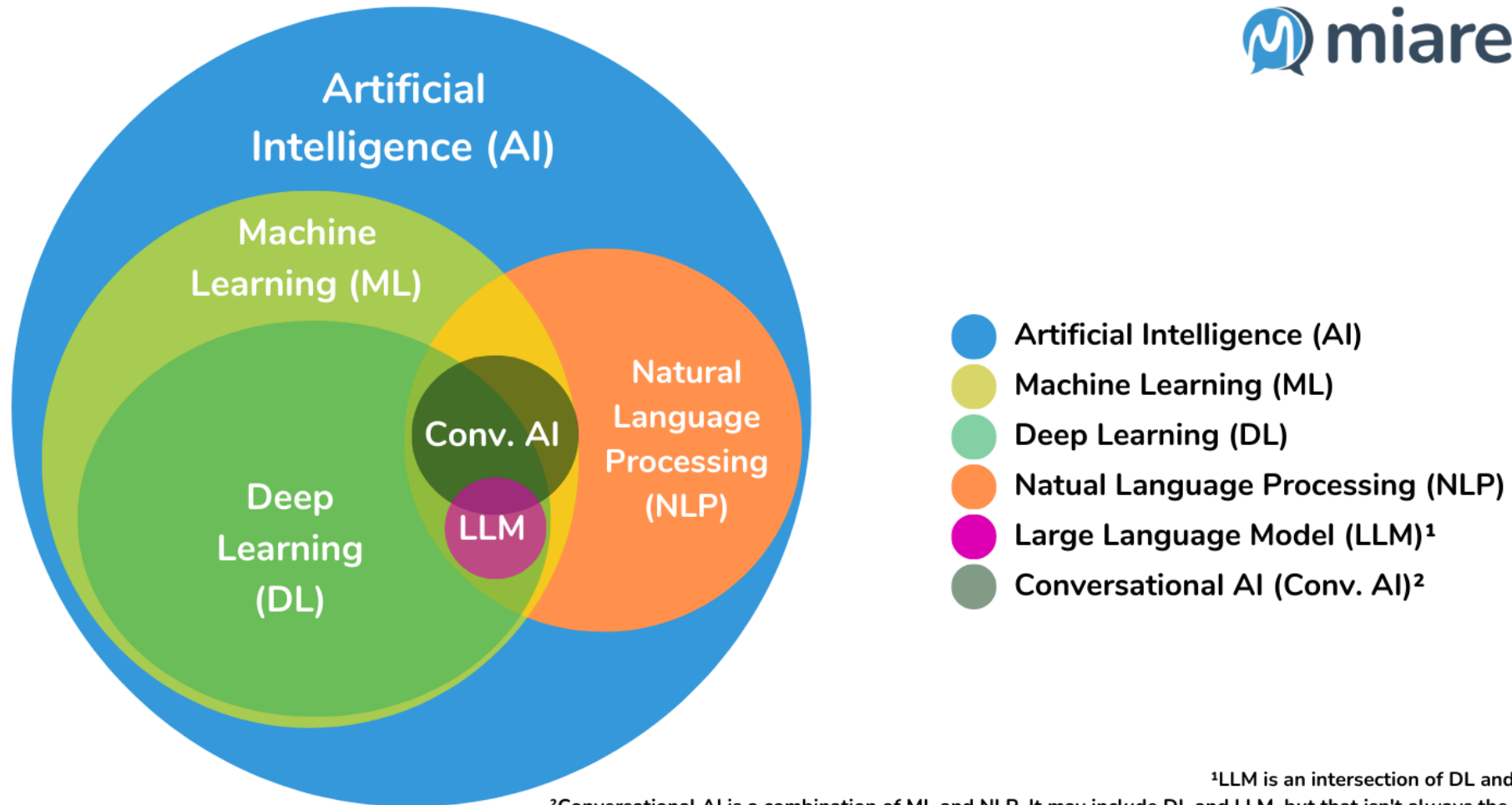
Codes for prediction methods

Codes for BERTopic Modeling

Some of the plots do not appear in
GITHUB (produced by plotly)

<https://github.com/ali-unlu/GMU>

What does AI bring new in Social Science Research?



¹LLM is an intersection of DL and NLP

²Conversational AI is a combination of ML and NLP. It may include DL and LLM, but that isn't always the case.

Handling Complex & High-Dimensional Data

Machine Learning (Inductive):

- Learns patterns **directly from data**.
- AI processes **unstructured data** (e.g., social media, images, survey responses).
- Uncover **patterns** and interactions that might be missed by conventional techniques.

Statistics (Deductive):

- Begins with **theory or hypothesis** and tests it using data.
- Relies on **pre-defined models** and **assumptions** (e.g., relationships between variables).
- Works with **structured, tabular data** (e.g., surveys).



Discovering Non-Linear Relationships

Machine Learning:

- Flexible and assumption-free.
- Handles **complex, non-linear relationships** effectively.
- Automates feature selection and pattern recognition.

Traditional Statistics:

- Relies on **strict assumptions** (e.g., normality, linearity).
- Works best for **simple, structured relationships**.
- Requires **manual selection of variables** and preprocessing.



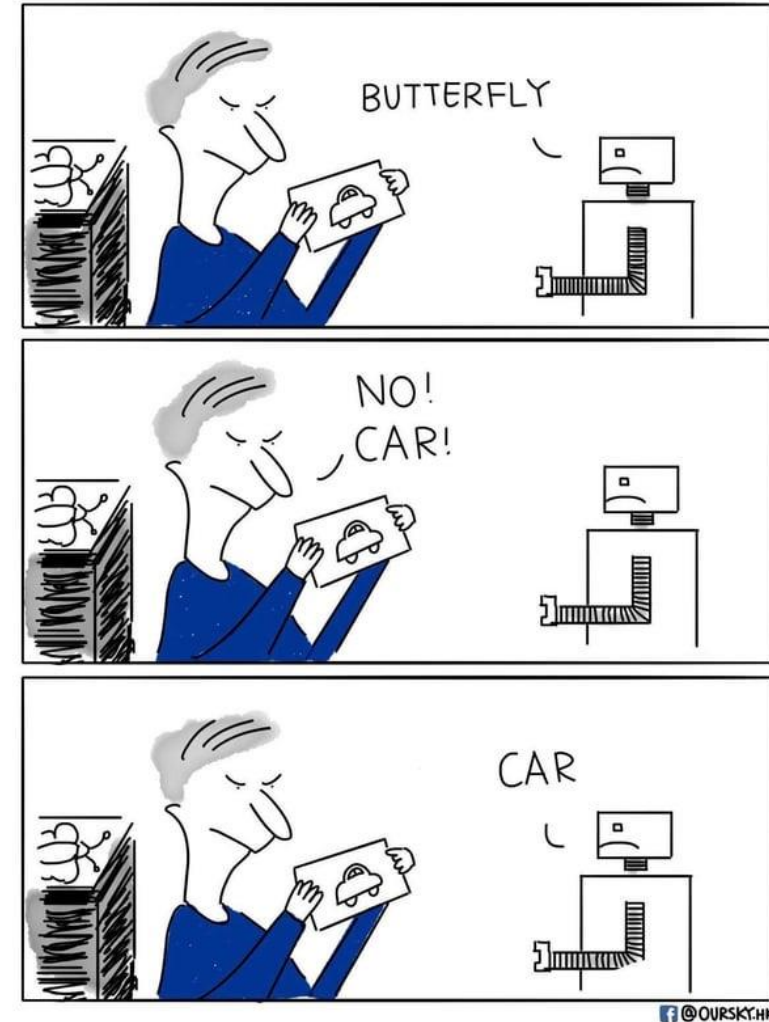
Predictive Accuracy and Scalability

Machine Learning:

- **Optimized for prediction** accuracy.
- Sacrifices **interpretability** for performance.
- **Scale** to massive datasets

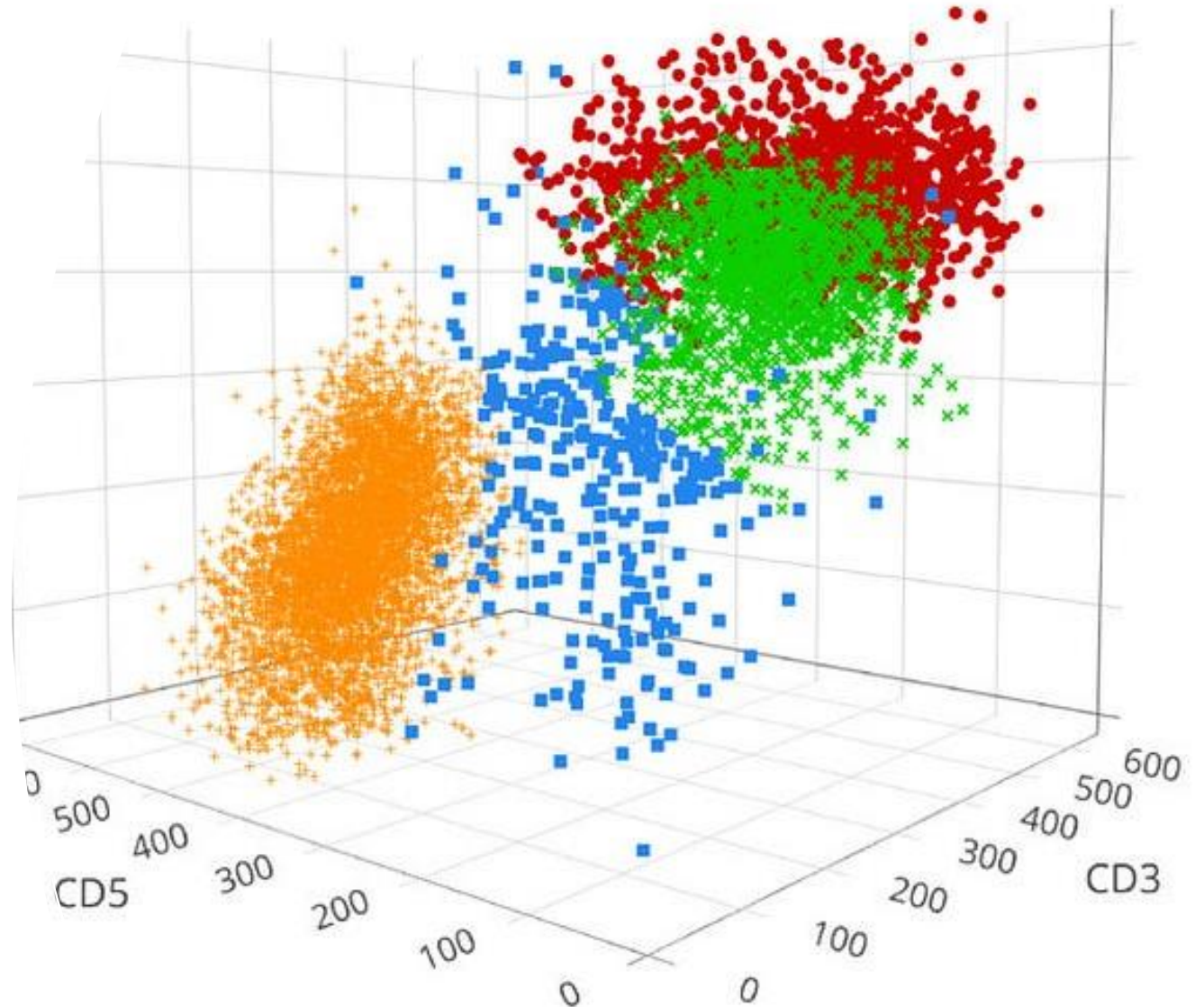
Traditional Statistics:

- Focuses on **explaining relationships** between variables.
- Suitable for research on **causal mechanisms** and hypothesis testing.
- Ideal for studies where understanding **impact** is critical (e.g., how income affects voting behavior).



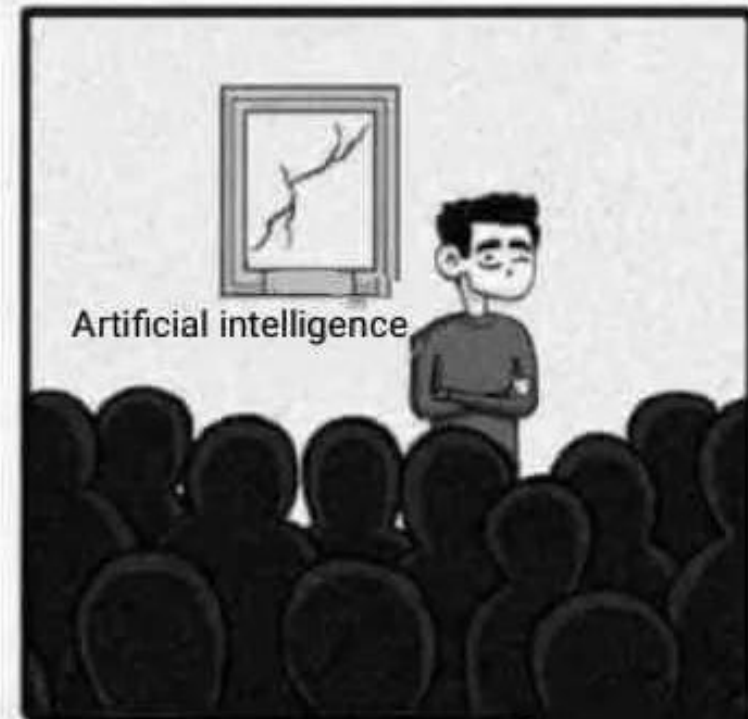
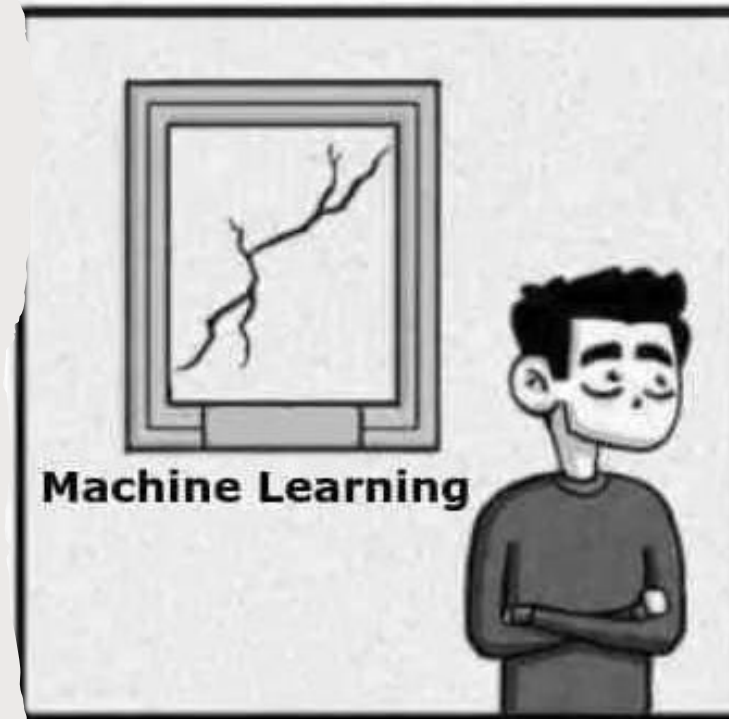
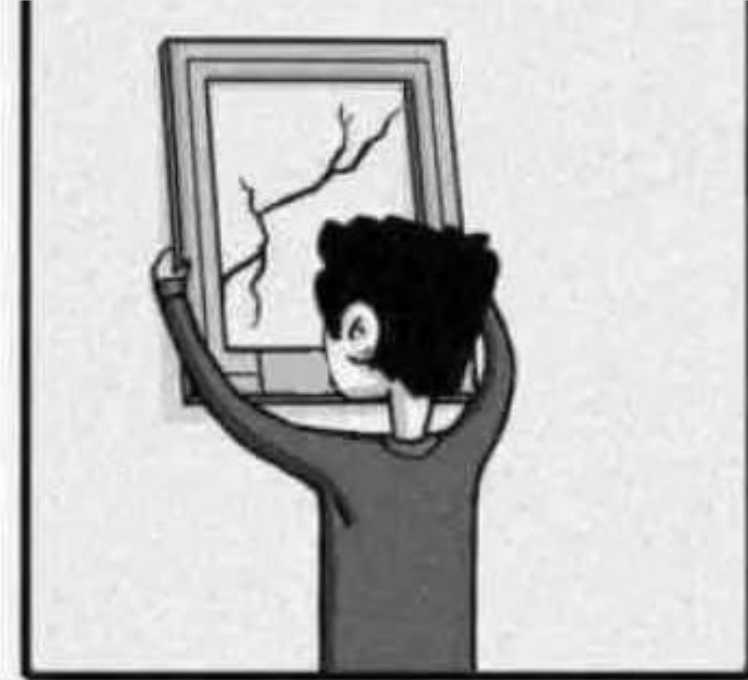
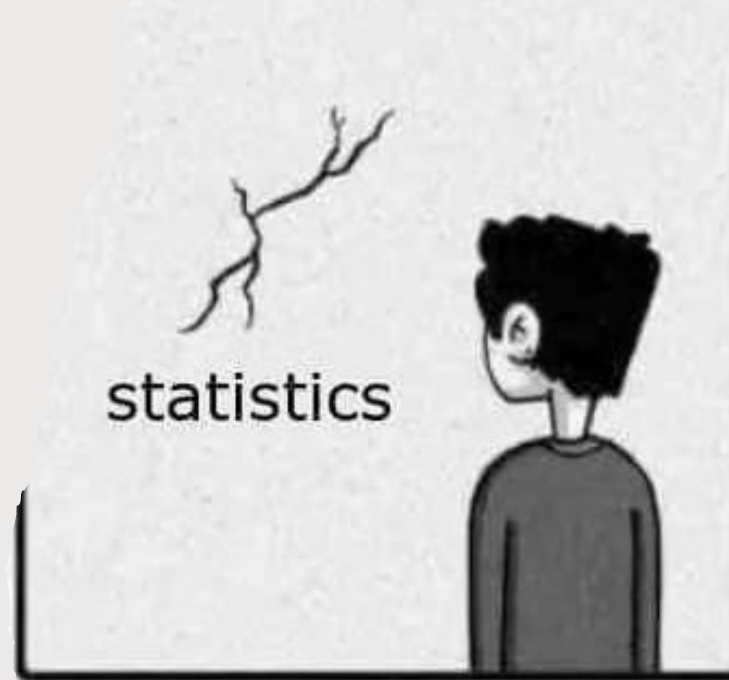
Facilitating Exploratory Analysis & Hypothesis Generation

- Unsupervised techniques (e.g., clustering, topic modeling) reveal latent structures that can inform new theories.



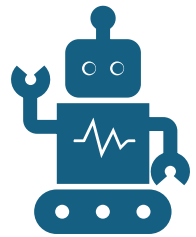
Complementing Traditional Statistical Approaches

- AI serves as a valuable complement, identifying relationships that can then be rigorously tested for causation using conventional statistics.

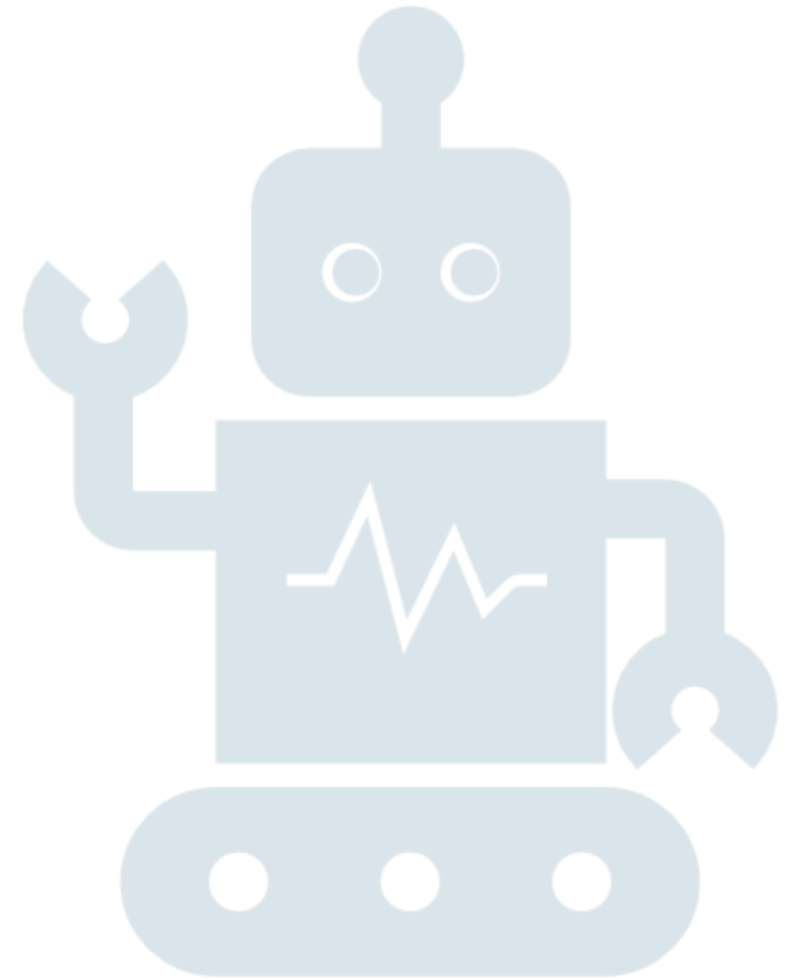


Basics of Predictive Modeling



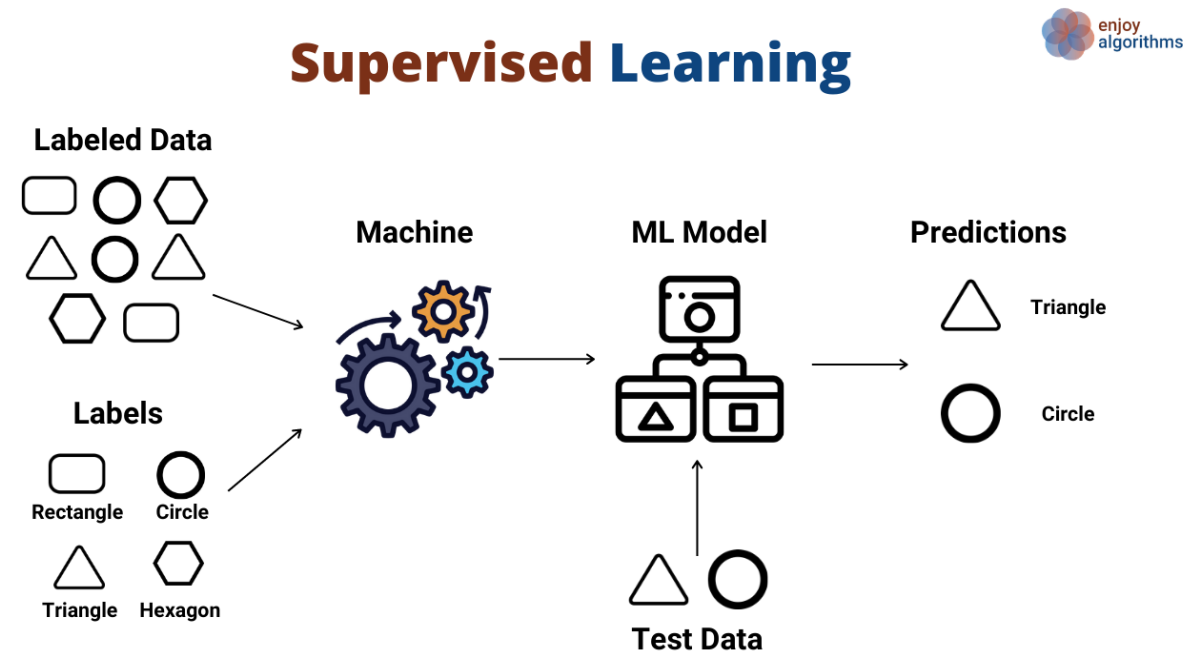


What does
prediction mean for
Machine Learning
and Social Science?

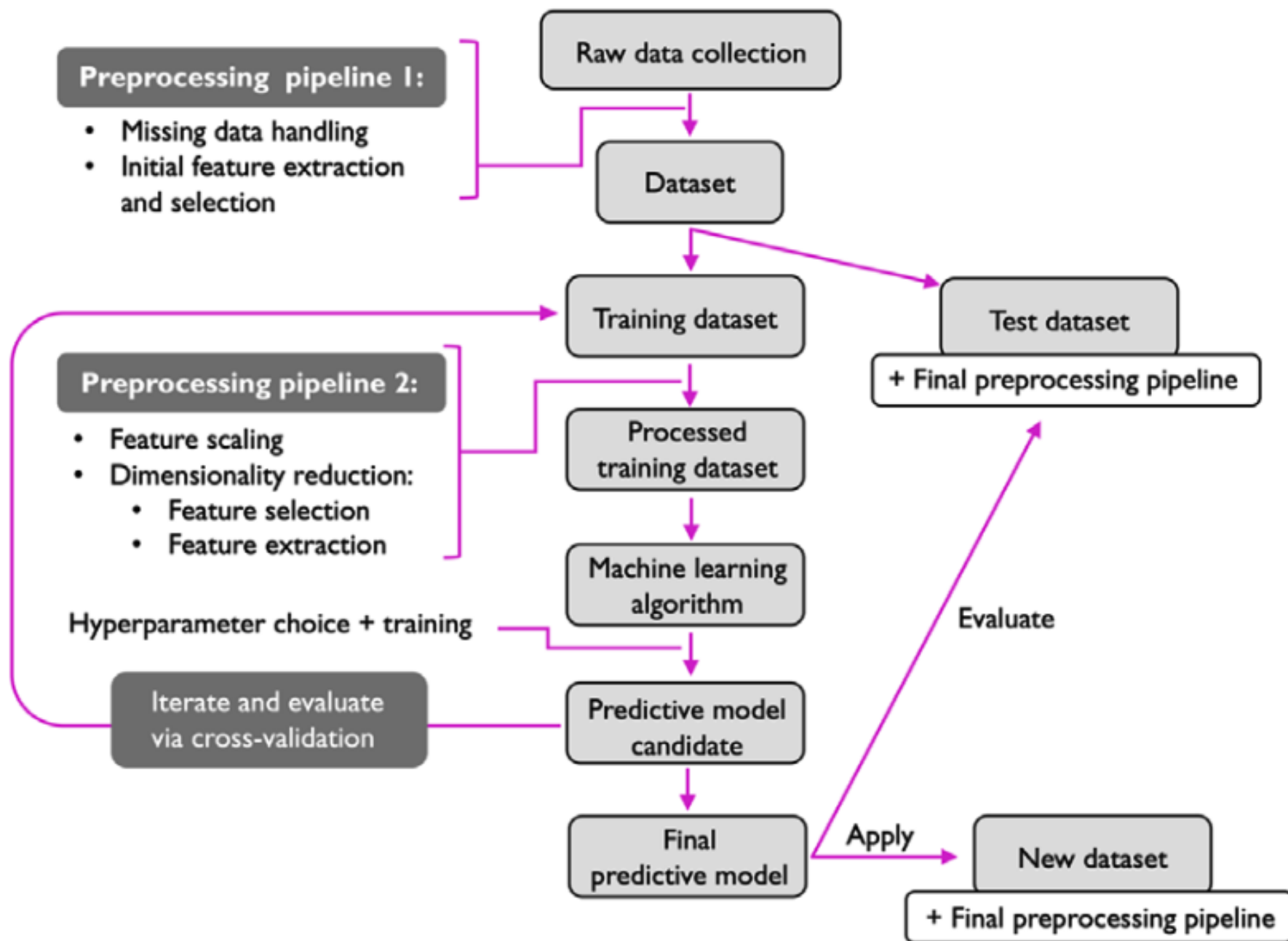


Prediction for AI and Social Science

- Machine Learning:
 - Mathematical tool to predict future or unseen cases
 - Relies on probabilistic reasoning
 - Performance Metrics (e.g., accuracy, precision, recall, F1-score, or mean squared error)
 - Algorithmic and Automated
- Social Science
 - Seek to explain causal mechanisms underlying social phenomena
 - Contextual and Interpretative
 - Policy and Practical Implications



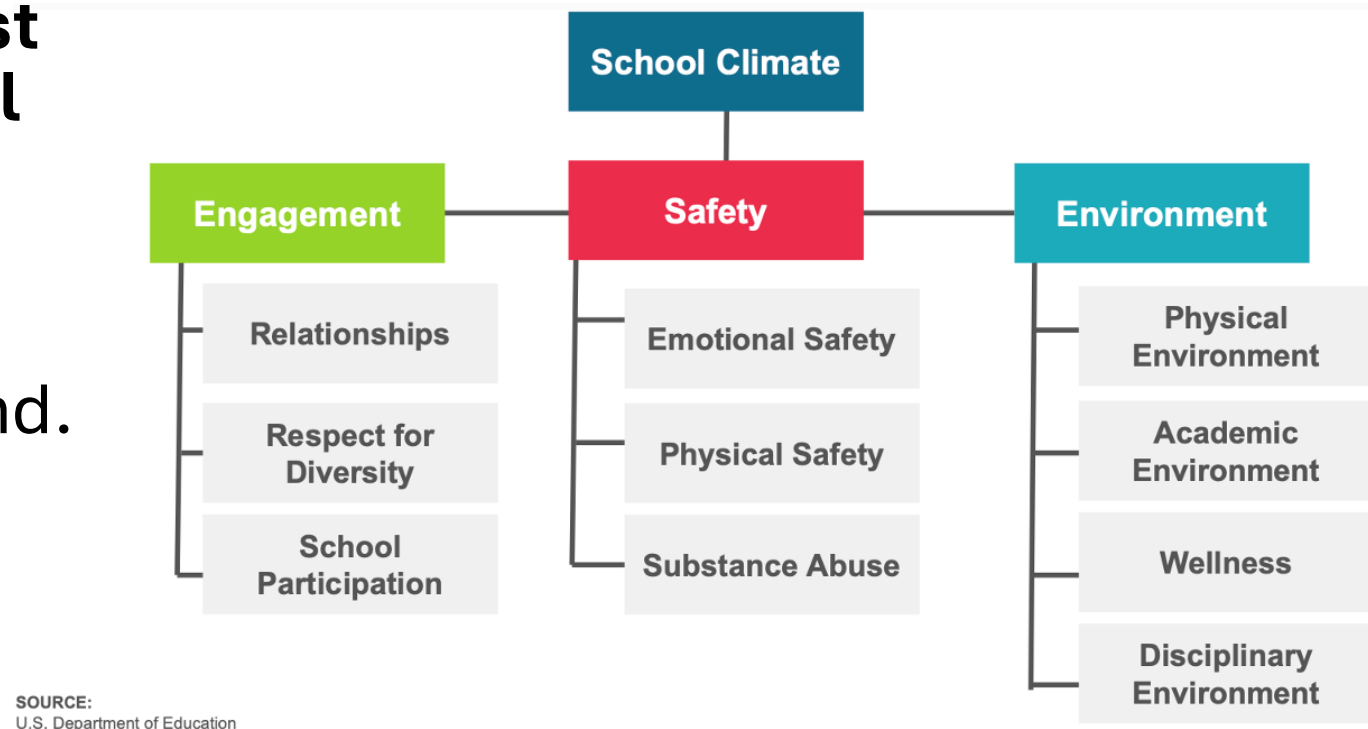
Workflow of predictive modeling



Case Study: School Climate and substance use

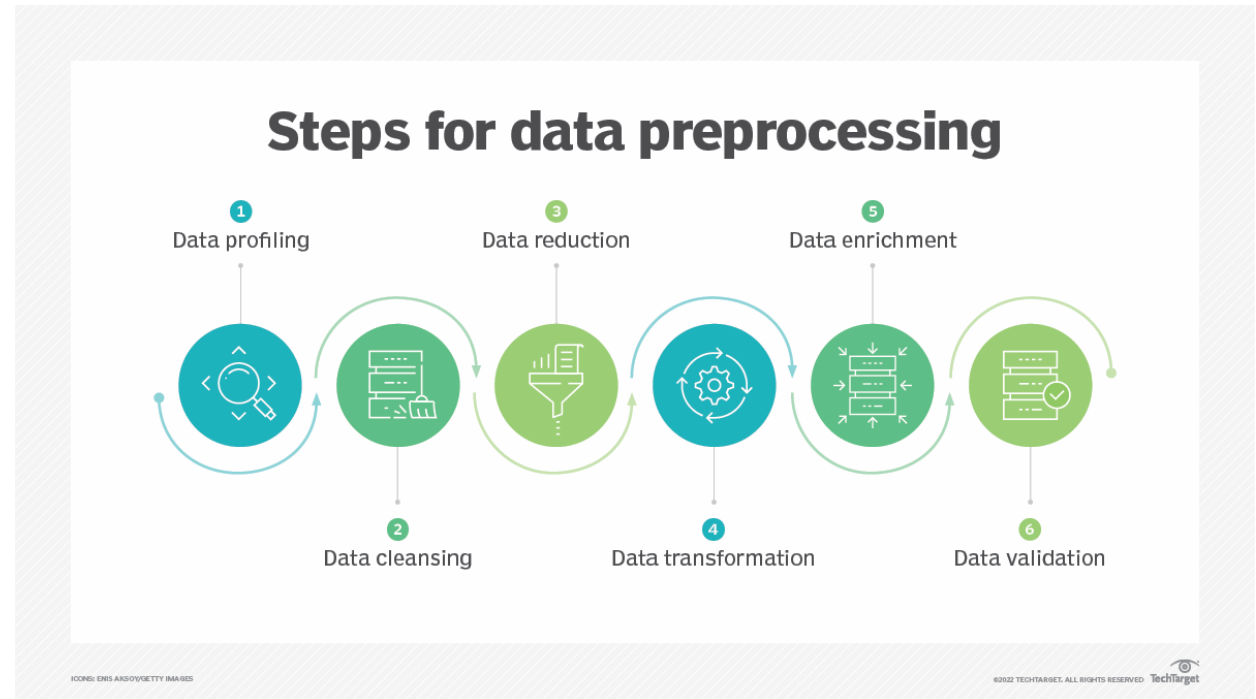
Which School Climate Indicators have the strongest predictive impact on alcohol and cannabis use?

- **Sample:** 69,513 students enrolled in 114 middle and high schools across Maryland.
- **Variables Analyzed:** 154 independent variables (excluding substance-use-related variables)



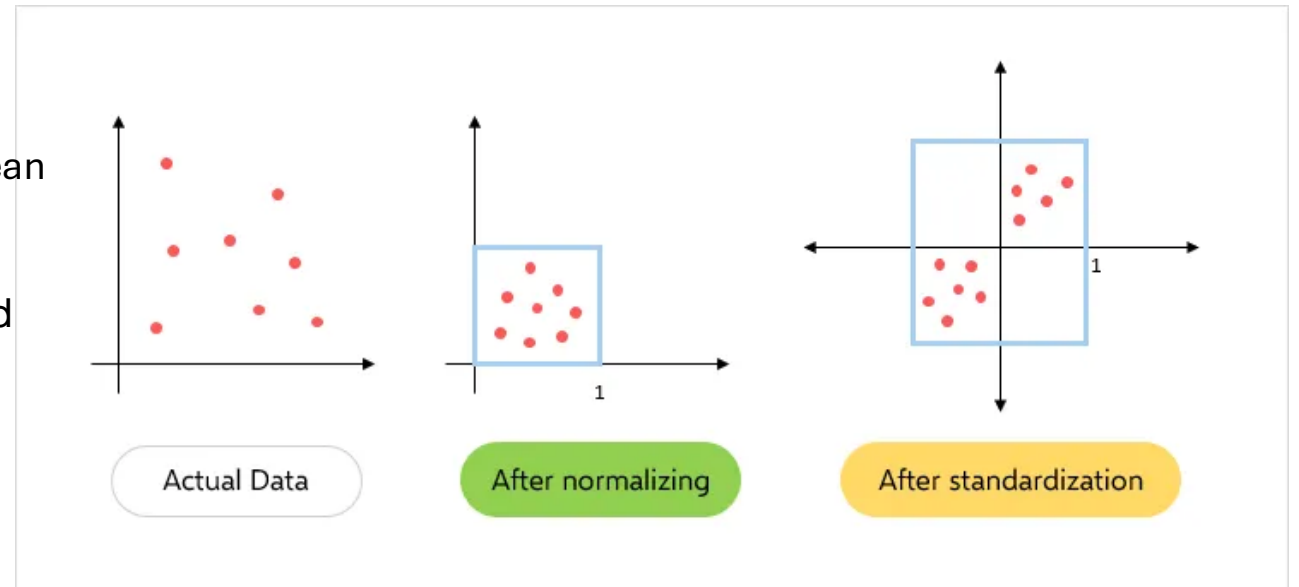
Data Preprocessing

- Data profiling.
 - Sweetviz python package
- Identify and sort out missing data.
 - decide whether it is better to discard records with missing fields, ignore them or fill them in with a probable value.
- Reduce noisy data
- Identify and remove duplicates
- Data transformation.



Data Normalization & Standardization

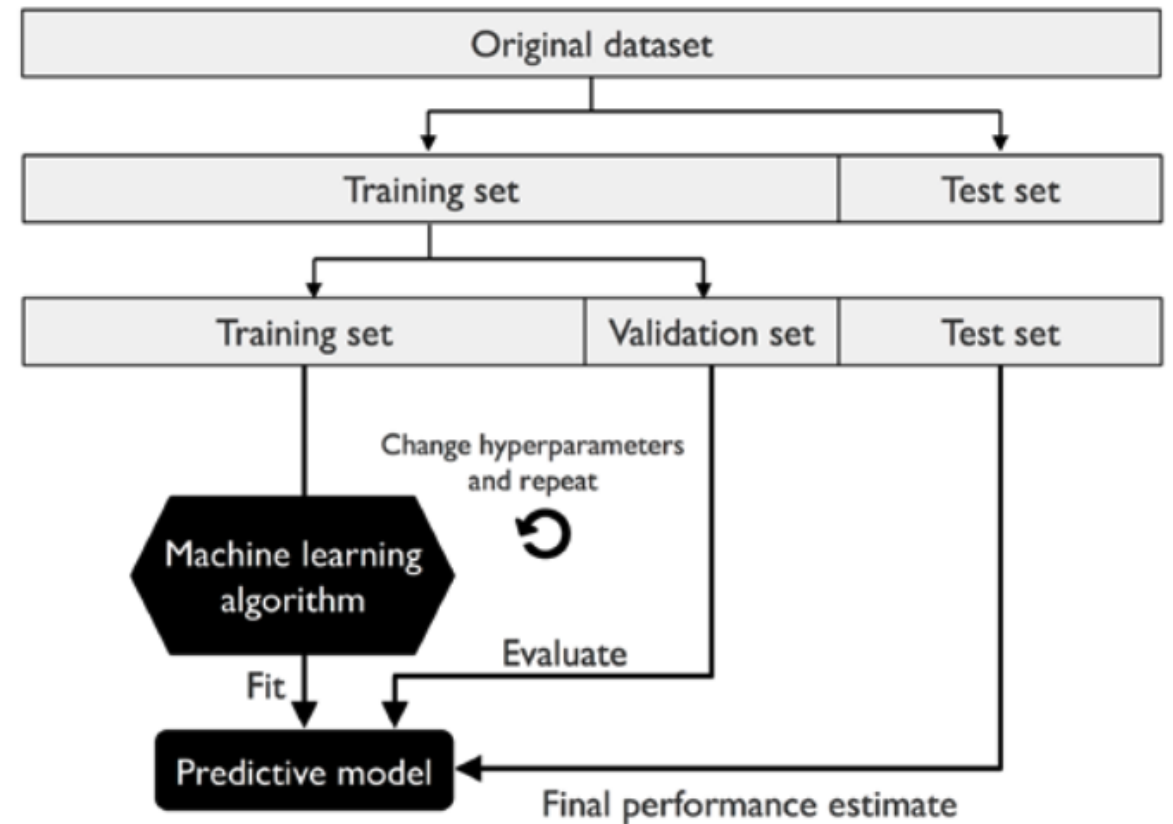
- **Normalization:** make the data homogenous over all records and fields.
- **Standardization:** the process of placing dissimilar features on the same scale
 - rescaling the attributes in such a way that their mean is 0 and standard deviation becomes 1
- Zero mean normalization or standardization
 - centers the data with a mean of 0 and a standard deviation of 1
 - StandardScaler()
- Decimal Scaling Normalization
 - normalizes by moving the decimal point of values of the data
 - values typically in the range of -1 to 1
- Min-Max Normalization
 - Rescales features to a specific range, typically 0 to 1
 - MinMaxScaler()



How can we validate ML results?

Train, Validation, and Test Split

- **Purpose:** Estimate model performance and generalization ability.
- **Process:**
 - **Training Set:** Model learn the task.
 - **Validation Set:** Which model is the best?
 - **Test Set:** How good is this model truly?



What does validation ensure?

- **Avoid Overfitting:** Prevents the test data from influencing model selection.
- **Better Generalization:** Ensures unbiased performance estimation.

<i>Prediction</i>	<i>Actual Values (ground truth)</i>	<i>Th = 0.2 >0.2, =1 <=0.2, =0</i>	<i>Th = 0.3 >0.3, =1 <=0.3, =0</i>	<i>Th = 0.4 >0.4, =1 <=0.4, =0</i>
0.6	1	1	1	1
0.5	1	1	1	1
0.4	0	1	1	0
0.3	0	1	0	0
0.2	0	0	0	0
TP		2	2	2
FP		2	1	0
FN		0	0	0
TN		1	2	3
tpr		2/2 = 1	2/2 = 1	2/2 = 1
fpr		2/3	1/3	0/3 = 0

(Th = Threshold)

Dimension Reduction & Feature Selection

Dimension reduction would not be feasible for many social science research, but feature selection

Feature Selection

- Identifies and retains the most **relevant features** from the dataset.
- **Eliminates redundant or irrelevant features** to simplify the model.
- Focuses on **preserving original features** without transformations.
- Filter wrapper, embedded models, etc

Feature Extraction

- **Transforms or combines existing features** into **new features** that capture meaningful information.
- Focuses on **dimensionality reduction** while retaining critical patterns.
- Reduces complexity by creating **compact representations** of data.
- Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), etc.

All Features



Feature Selection

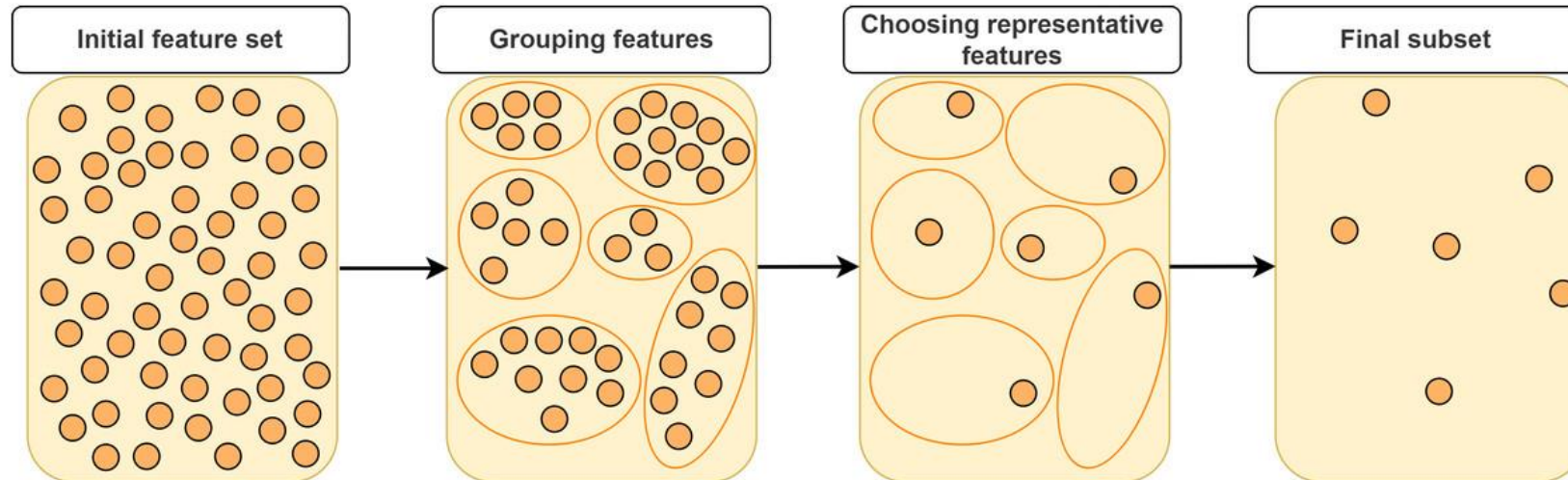


Final Features



Challenge of Feature Selection Methods

- Different feature selection methods yield varying top features yet similar prediction scores.
 - Which algorithm to rely on?
 - Which results to take further steps?
 - How to justify your choice?



Key Considerations Beyond Accuracy

- **Stability:** Consistency of selected features across data splits
- **Interpretability:** Ease of explaining feature importance and alignment with domain knowledge
- **Computational Efficiency:** Resource and time demands of the method
- **Theoretical Alignment:** Compatibility with established theories and research objectives

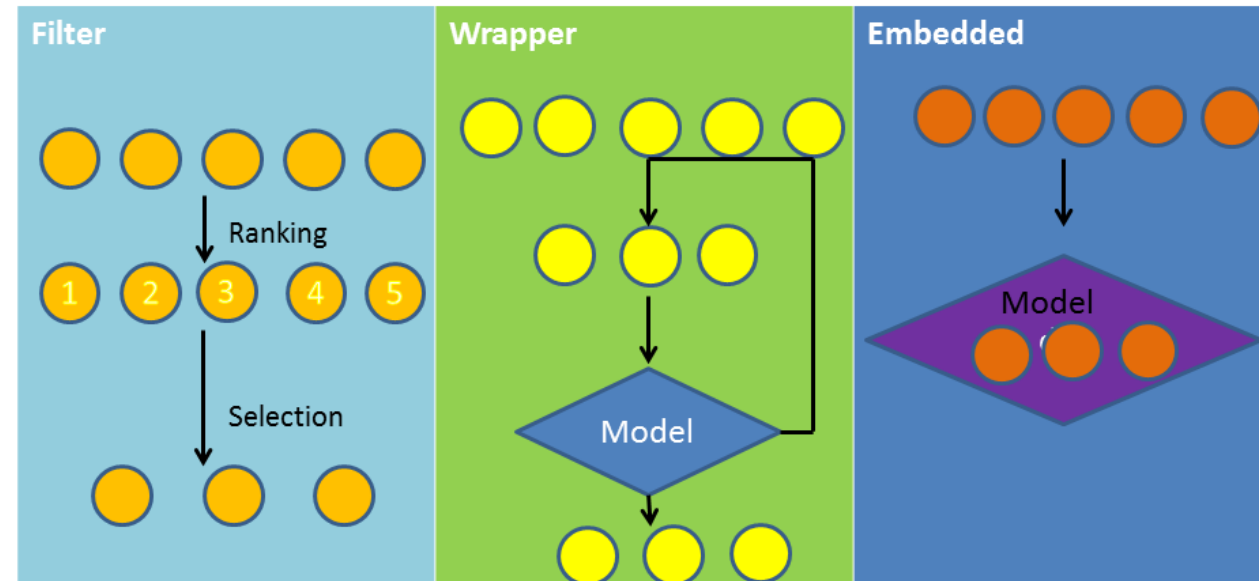


Table 4: Top 10 Features for Cannabis across Models																
	Variable	RFE RF	RFE GBC	Embedded Lasso LR	Embedded RF	SFS KNN	SFS LR	SBS RF	SBS LR	Information Gain	K Best	LSTM	BILSTM	Recursive LSTM	Boruta XGB	Boruta RF
1	Age	✓			✓									✓	✓	✓
2	Snus	✓	✓	✓						✓	✓		✓			
3	E-cigarette	✓		✓	✓				✓	✓				✓		✓
4	Health risk (regular use)	✓			✓					✓	✓				✓	
5	Health Risk (experimental use)	✓			✓					✓	✓		✓	✓	✓	
6	No punishment	✓	✓													
7	Punishment growing cannabis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
8	Drug user friend	✓			✓	✓	✓			✓	✓	✓	✓	✓	✓	✓
9	Free drug offer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓
10	Purchase drug offer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		
11	Cigarette		✓		✓			✓				✓	✓		✓	✓
12	Nicotine pouch		✓												✓	
13	Alcohol use		✓									✓				✓
14	Punishment picking mushroom		✓			✓							✓			
15	Punishment heroin use		✓	✓							✓					
16	Punishment cannabis use				✓		✓		✓	✓	✓					
17	Drug testing at work (HR)			✓		✓			✓	✓						✓
18	Alcohol related health problem			✓												
19	Drug consumption room (HR)								✓			✓	✓	✓		✓
20	Punishment mailing cocaine			✓							✓	✓				
21	Drug related health problem			✓										✓		
22	Police purchase (HR)															✓
23	Substitution (HR)												✓	✓		
24	Cannabis accessibility									✓		✓			✓	✓
25	Intoxicant accessibility														✓	
26	Antidepressant prescription												✓			
27	Drug problem at country level							✓				✓	✓		✓	
28	Drug problem at neighborhood level													✓		
29	Buprenorphine prescription						✓									
30	Pain killer prescription						✓	✓								
31	Current health status											✓				
32	City type (Capital region)													✓		
33	Marriage status (widow)					✓	✓									
34	Education (advance)								✓							
35	Gender (male)							✓								
36	Imprisonment (punishment)					✓										
37	City type (Other rural areas)						✓									
38	City type (A rural settlement or agglomeration)						✓									
39	City type (A city of 50,000 to 100,000 inhabitants)											✓				

Selected Features

0	U031SV24R	Have you ever belonged to a gang?	Common
1	U031SV23R2	During the last month, how many days of school have you missed because you skipped or "cut"?	Common
2	U031EI6R	I have threatened to hit or hurt someone	Common
3	U031SE6R	I do all my school work	Common
4	U031AE3R	My teachers encourage me to work hard in my classes	Common
5	U031EI1R	It is okay to hit someone if they hit me first	Common
6	U031WE13R	Get into a lot of trouble	Common
7	U031FI2R	If I do something bad at school, my parent(s) or guardian(s) hears about it	Common
8	U031FI3R	When I do something good at school, my parent(s) or guardian(s) usually hears about it	Common
9	U031WE26_8	I have never gambled, Have you ever:	Common
10	U031D6St	How old are you?	Common
11	U031AE1R	My teachers believe that I can do well in school	Common
12	U031EI7R	I do things without thinking	Common
13	U031SE1R	My teachers make me feel good about myself	Common
14	U031PS1R	The principal at my school cares about us students	Common
15	U031AE2R	I believe I can do well in school	Unique to Alcohol
16	U031SV19R	During the past 30 days, how often did you carry a weapon, such as a knife or gun, on school property?	Unique to Alcohol
17	U031D1	Gender	Unique to Alcohol
18	U031BU4R_0_4	In the past 30 days, how often have you bullied someone else?	Unique to Alcohol
19	U031CO20R	At this school, students and staff feel pride in this school	Unique to Alcohol
20	U031WE20R	Felt that difficulties were piling up so high that you could not overcome them	Unique to Cannabis
21	U031EI5R	I have trouble controlling my temper	Unique to Cannabis
22	U031CO8R	At this school, my teachers care about me	Unique to Cannabis
23	U031EI8R	I get mad easily	Unique to Cannabis
24	U031SE7R	I like this school	Unique to Cannabis

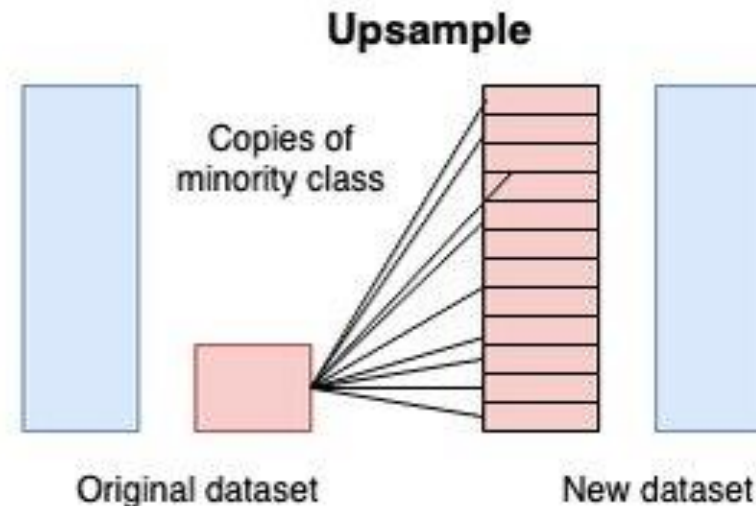
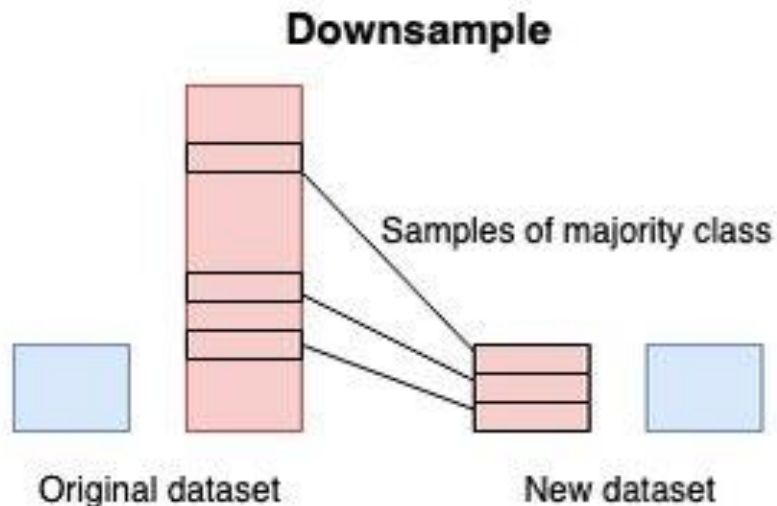
How does
data
imbalance
affect
results?



Imbalance data

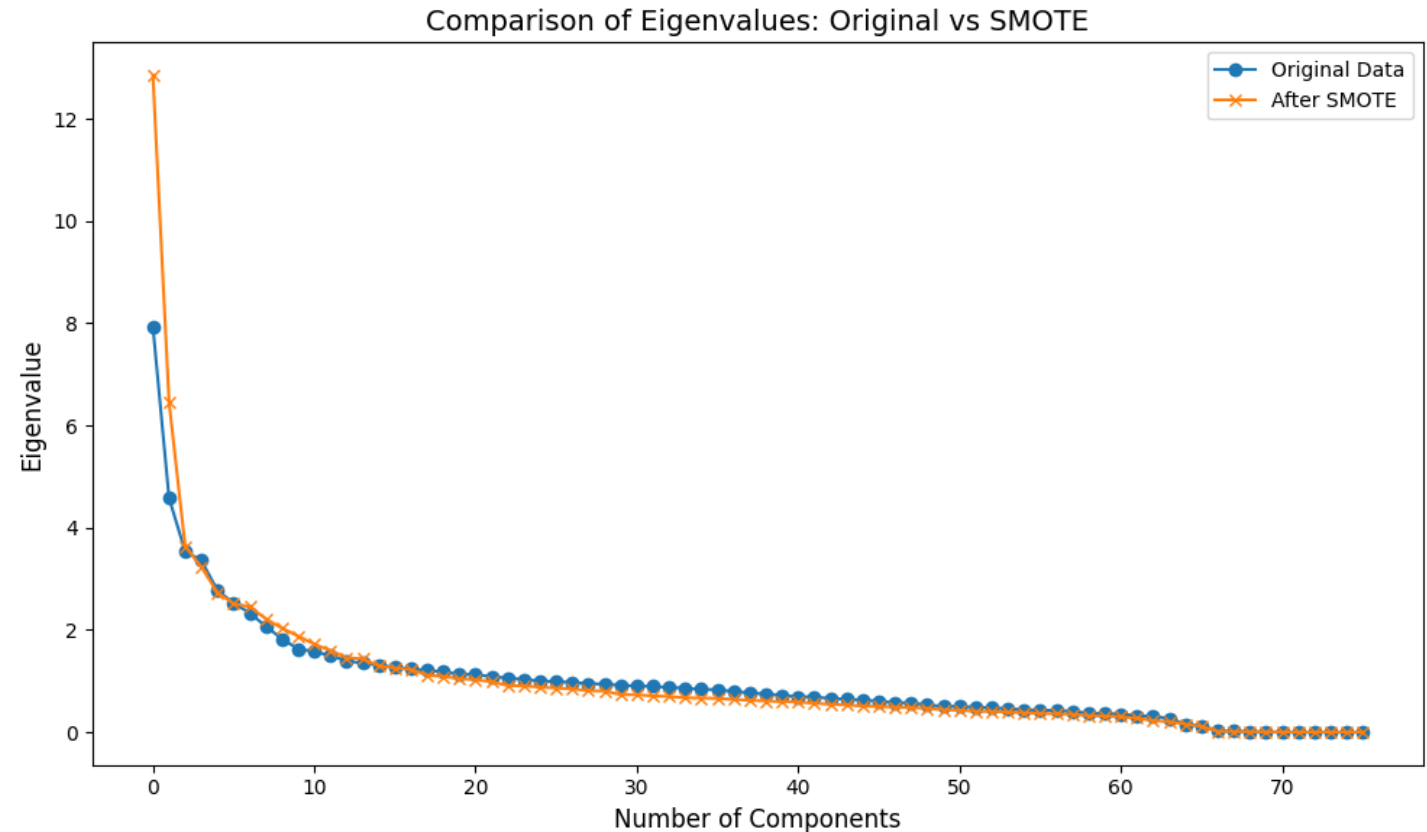
Techniques to Handle Imbalance:

- **Upsampling (Over-sampling):**
 - Increase minority class samples (e.g., **SMOTE**).
- **Downsampling (Under-sampling):**
 - Reduce majority class samples (e.g., **Random Sampling**).

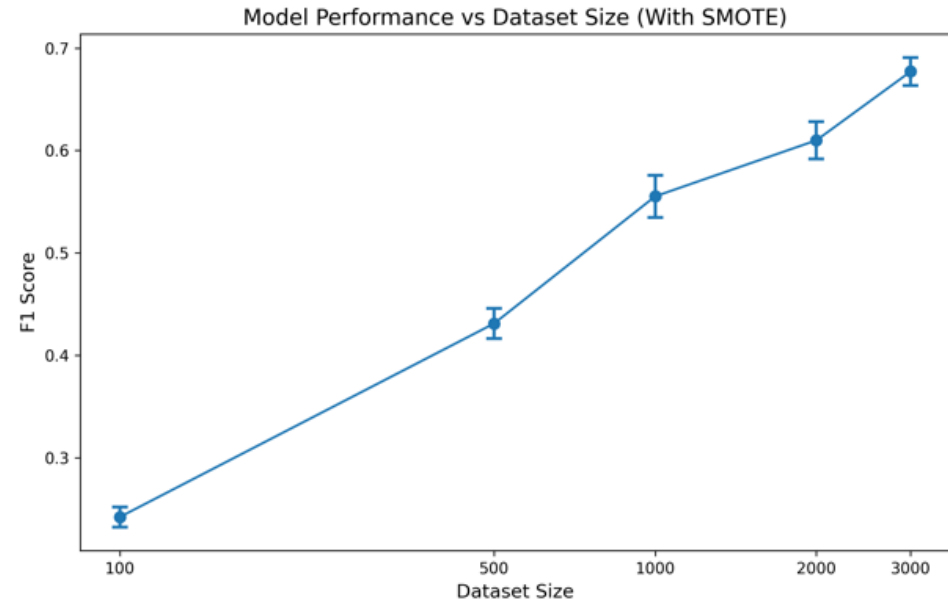
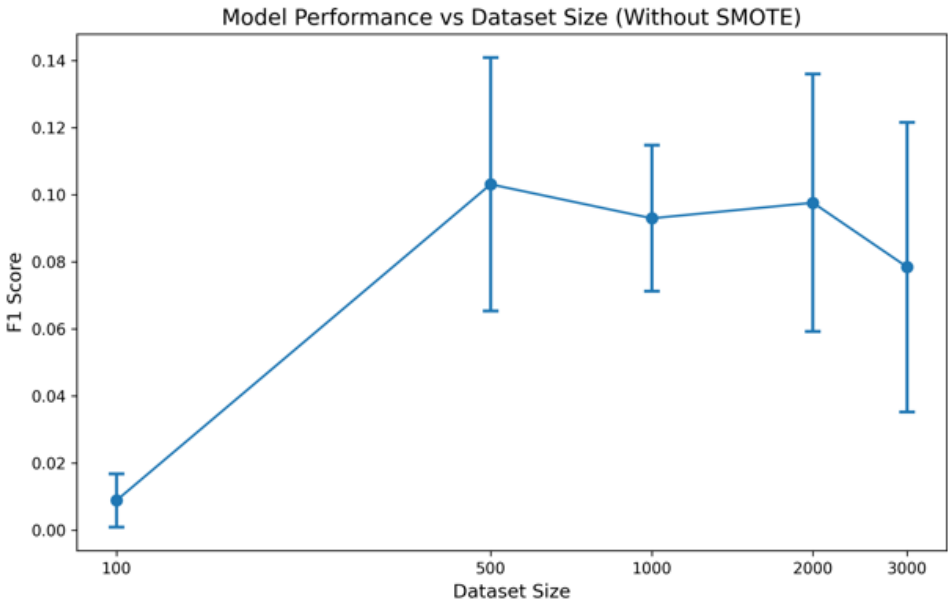
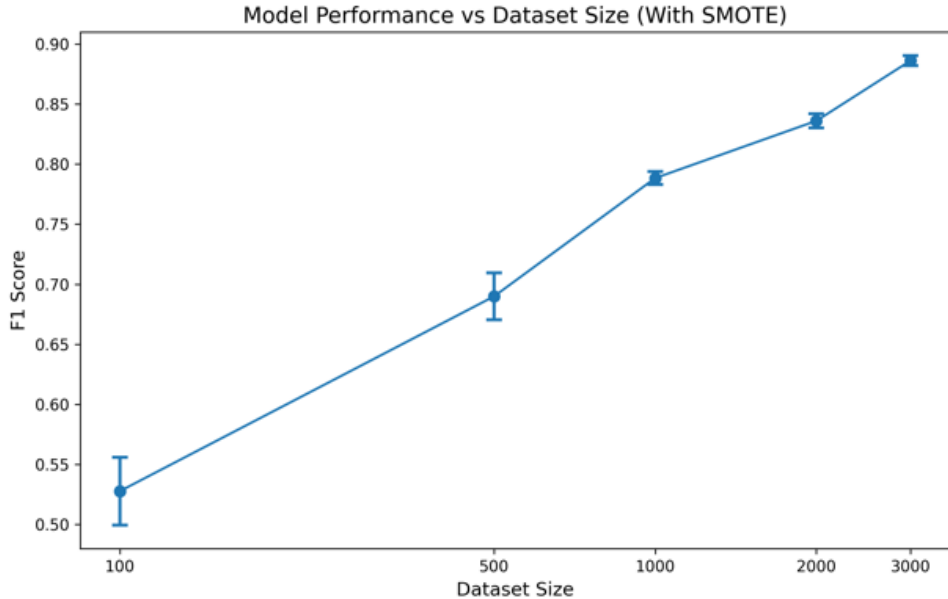
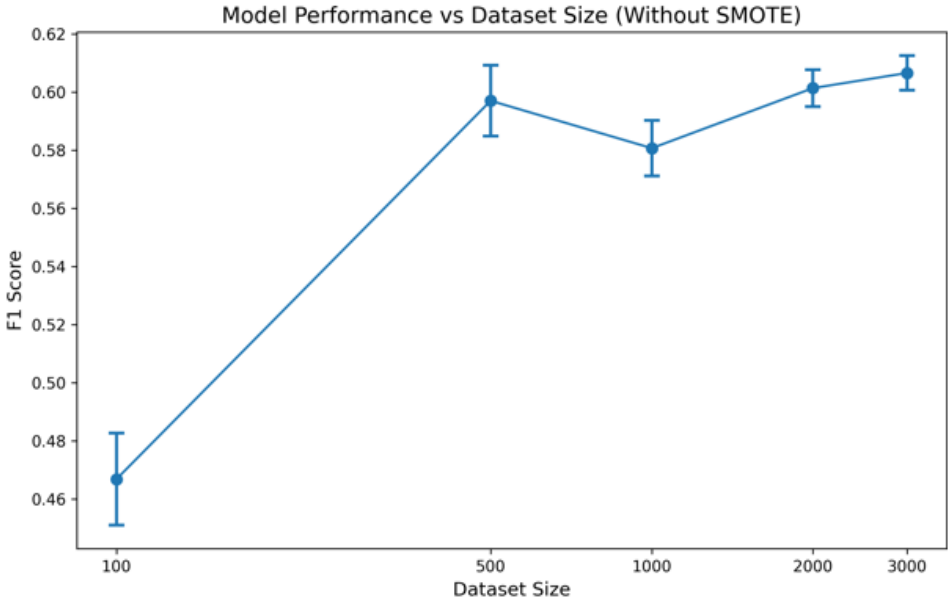


Key Considerations

- **Upsampling** reduces bias but risks **overfitting**.
- **Downsampling** saves memory but may cause **information loss**.

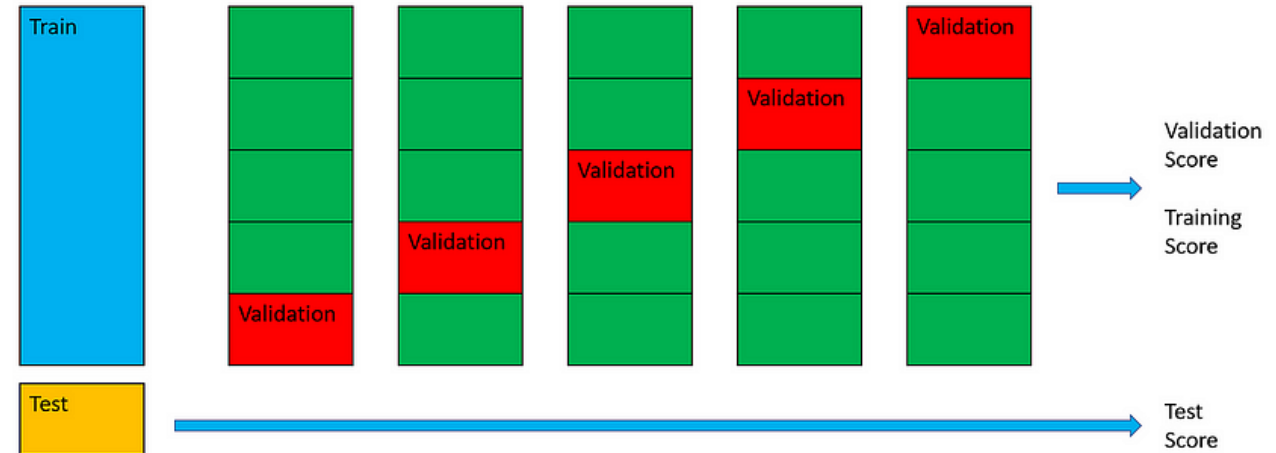


How does
SMOTE
affect
prediction?



Cross Validation

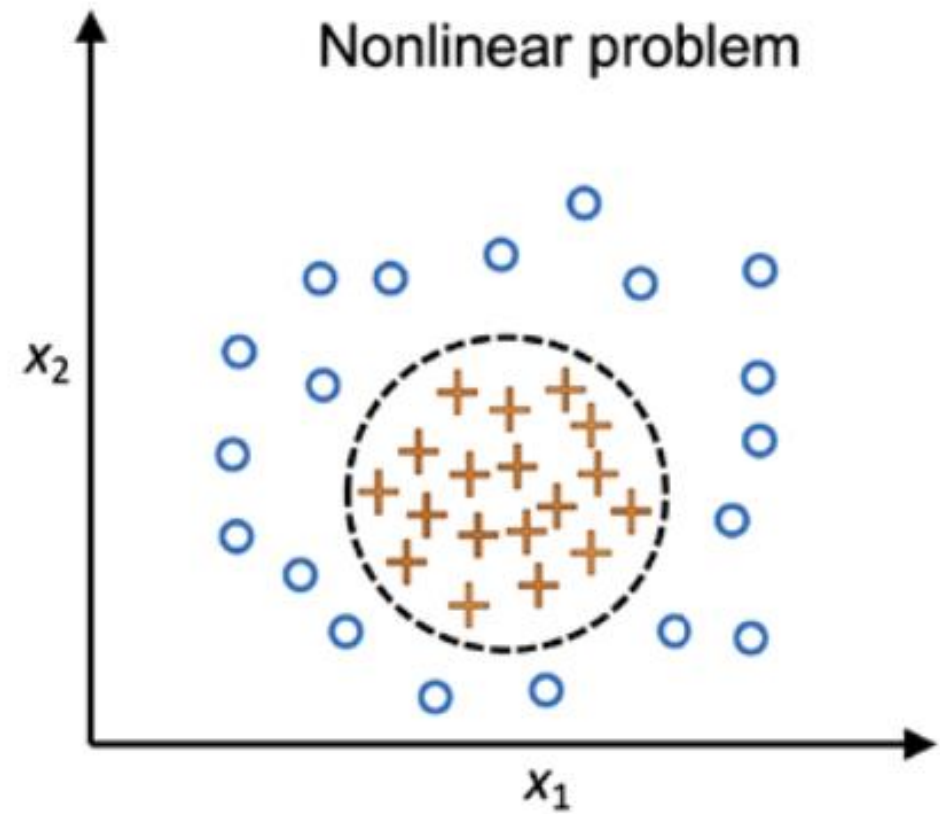
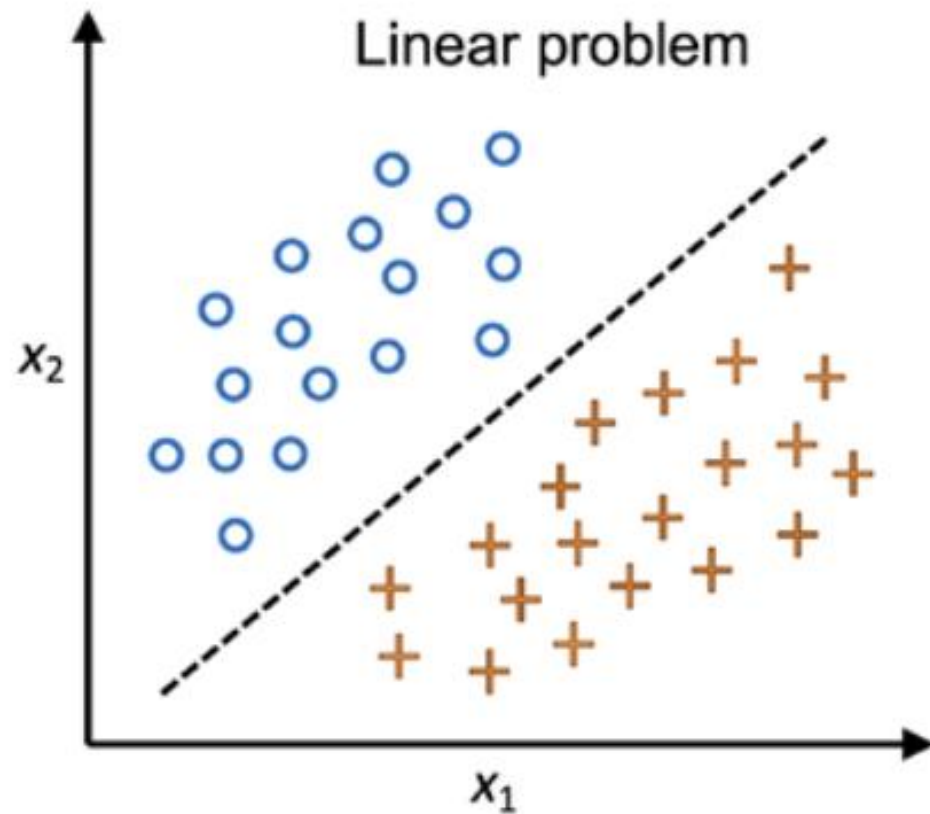
- K-fold Cross validation
 - Each split is used once for testing the model as it iterates through the training process.
- Leave-One-Out Cross-Validation (LOOCV)
 - One data point is used for testing and the entirety of the remaining data is used for training.
- Stratified K-fold Cross-Validation
 - Each subset will have an equal number of values for each class label



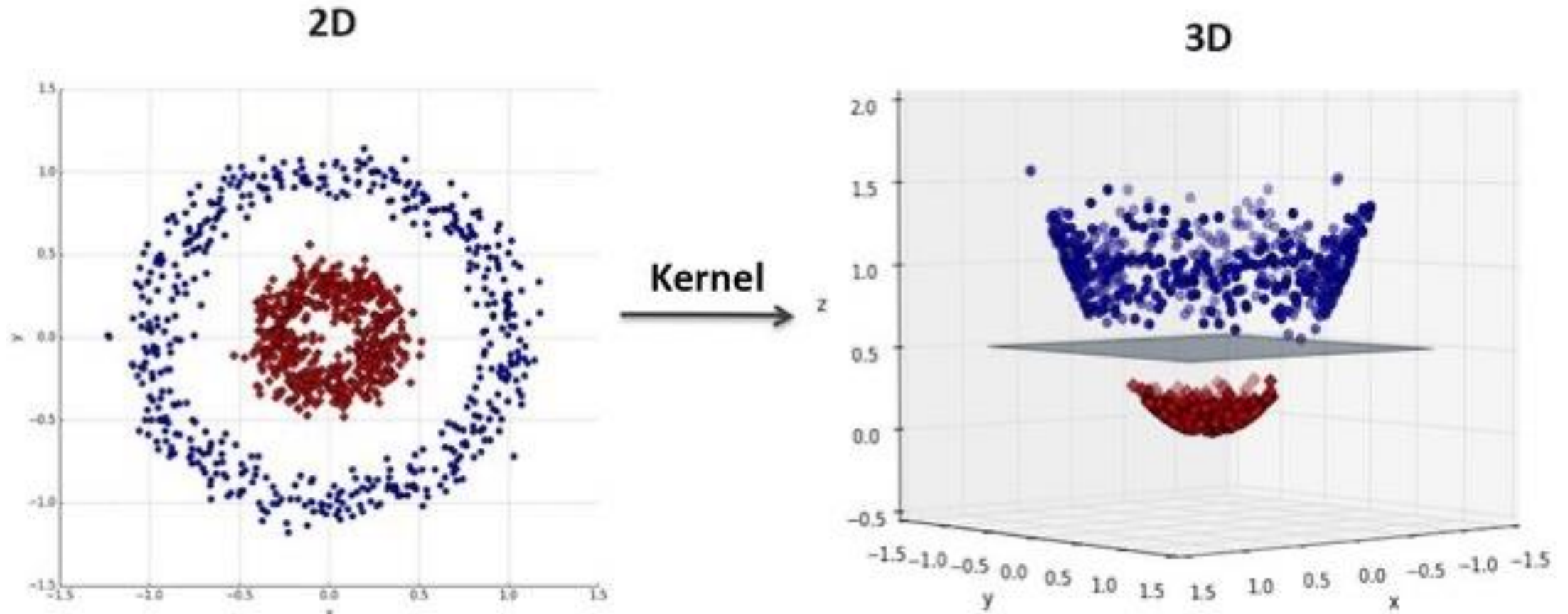
Algorithms



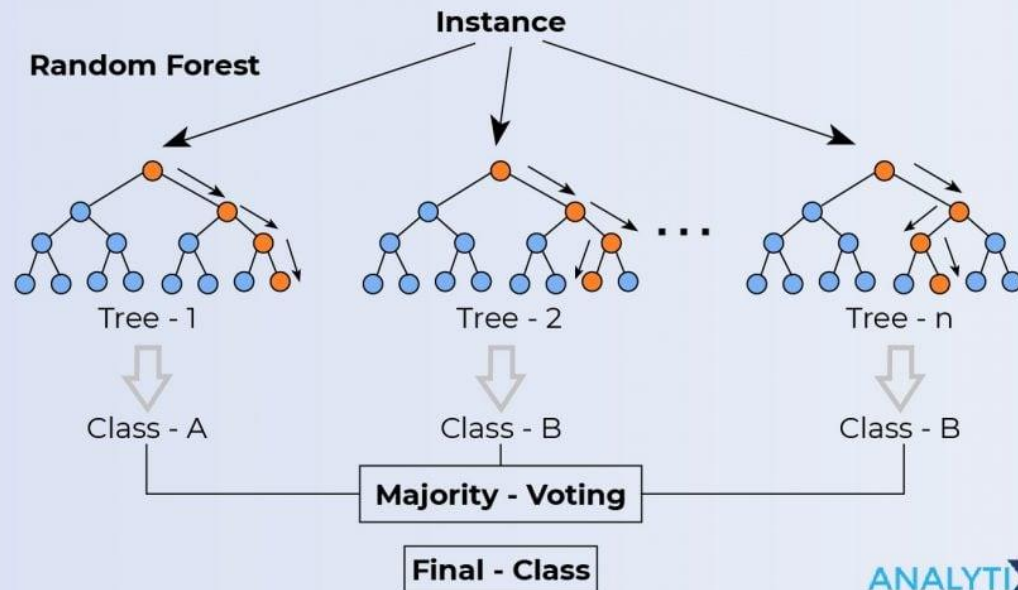
Nature of the Problem



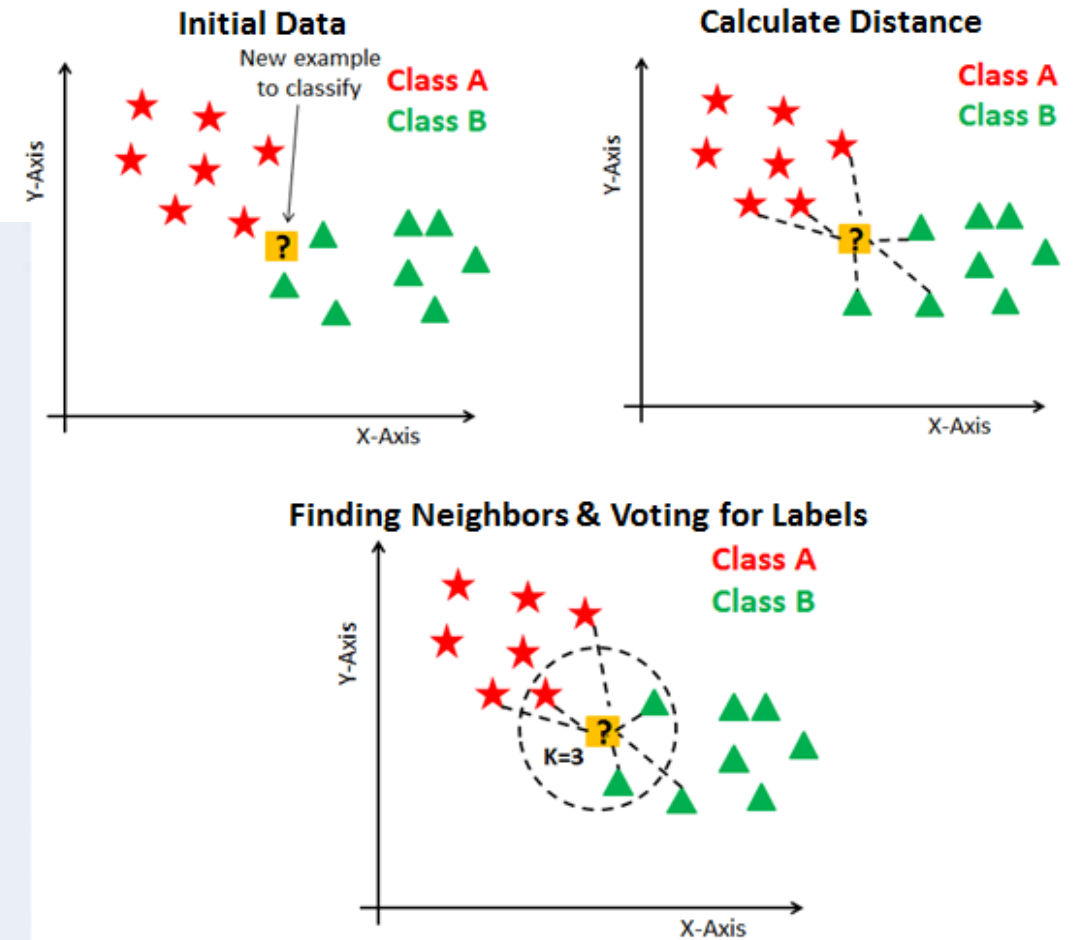
Non-linear classifier using Kernel trick



Random Forest Simplified



K-Nearest Neighbors



Hyper Tuning

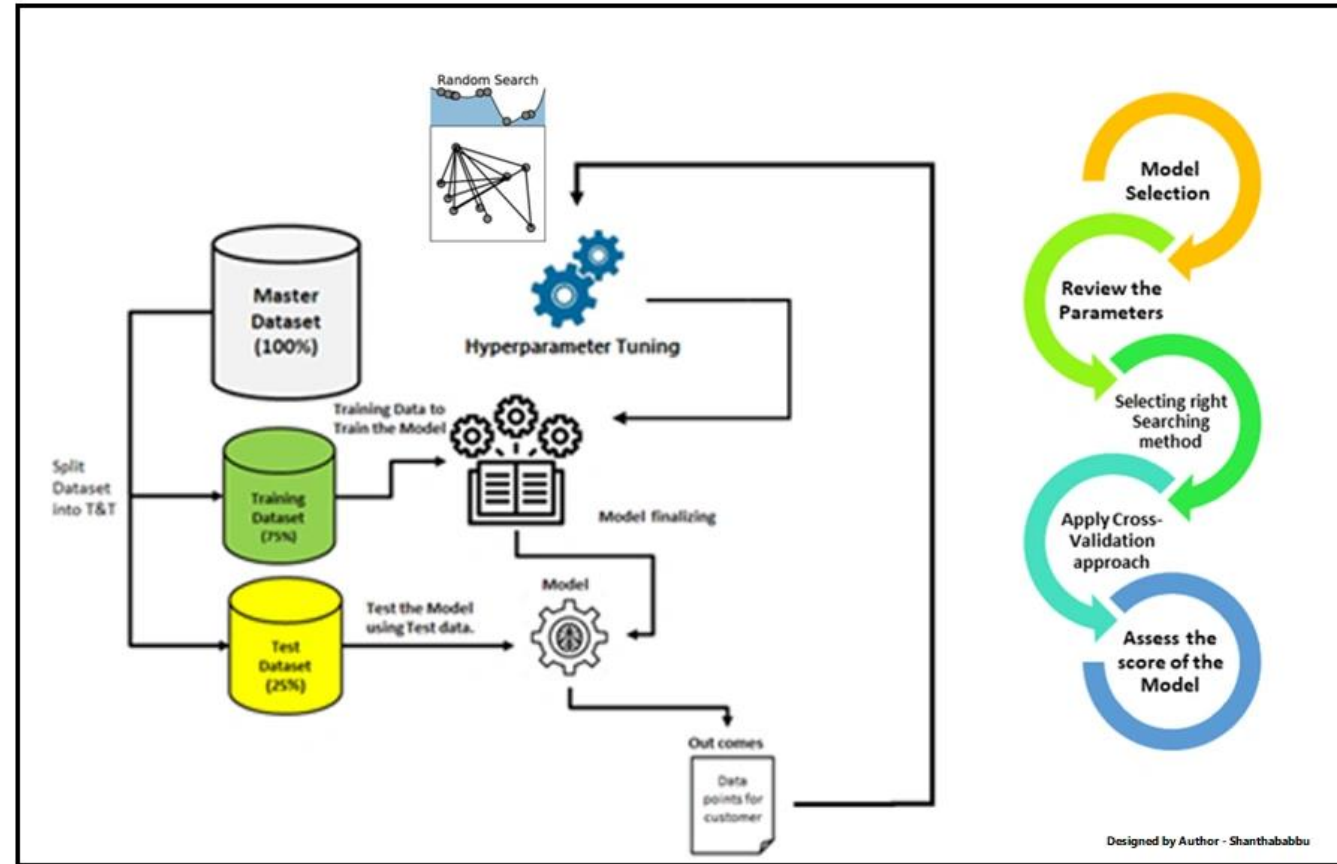
- Optimizes model parameters to improve performance.

Common Techniques:

- **Grid Search:**
 - Exhaustive search through predefined parameter values.
- **Random Search:**
 - Randomly samples parameters within a range.
- **Bayesian Optimization:**
 - Iteratively improves parameter selection based on prior evaluations.

Key Considerations:

- Balance **accuracy** with **computational cost**.
- Use **cross-validation** to validate results.



Determining best parameters

```
from sklearn.ensemble import ExtraTreesClassifier, RandomForestClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier
from sklearn import metrics
import os

# Define models
models = {
    "gbc": GradientBoostingClassifier(random_state=1),
    "xgboost": XGBClassifier(random_state=1, eval_metric="logloss"),
    "extra_trees": ExtraTreesClassifier(random_state=1),
    "random_forest": RandomForestClassifier(random_state=1),
}

# Define hyperparameter grids
param_grids = {
    "gbc": {
        "n_estimators": [100, 200, 250, 300],
        "learning_rate": [0.01, 0.1, 0.2],
        "max_depth": [3, 5, 10]
    },
    "xgboost": {
        "n_estimators": [100, 200, 250, 300],
        "learning_rate": [0.01, 0.1, 0.2],
        "max_depth": [3, 5, 10]
    },
    "extra_trees": {
        "n_estimators": [100, 200, 250, 300],
        "max_depth": [None, 10, 20],
        "min_samples_split": [2, 5],
        "min_samples_leaf": [1, 2]
    },
    "random_forest": {
        "n_estimators": [100, 200, 250, 300],
        "max_depth": [None, 10, 20],
        "min_samples_split": [2, 5],
        "min_samples_leaf": [1, 2]
    }
}
```



Hyperparameters

- ⚙️ n_layers = 3
n_neurons = 512
learning_rate = 0.1
- ⚙️ n_layers = 3
n_neurons = 1024
learning_rate = 0.01
- ⚙️ n_layers = 5
n_neurons = 256
learning_rate = 0.1



Parameters

- ⚙️ Weights optimization
- ⚙️ Weights optimization
- ⚙️ Weights optimization



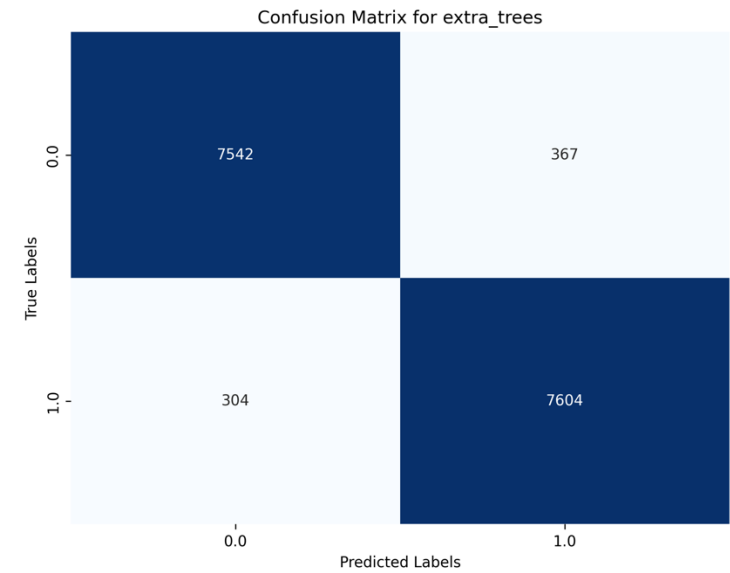
Score

85%

80%

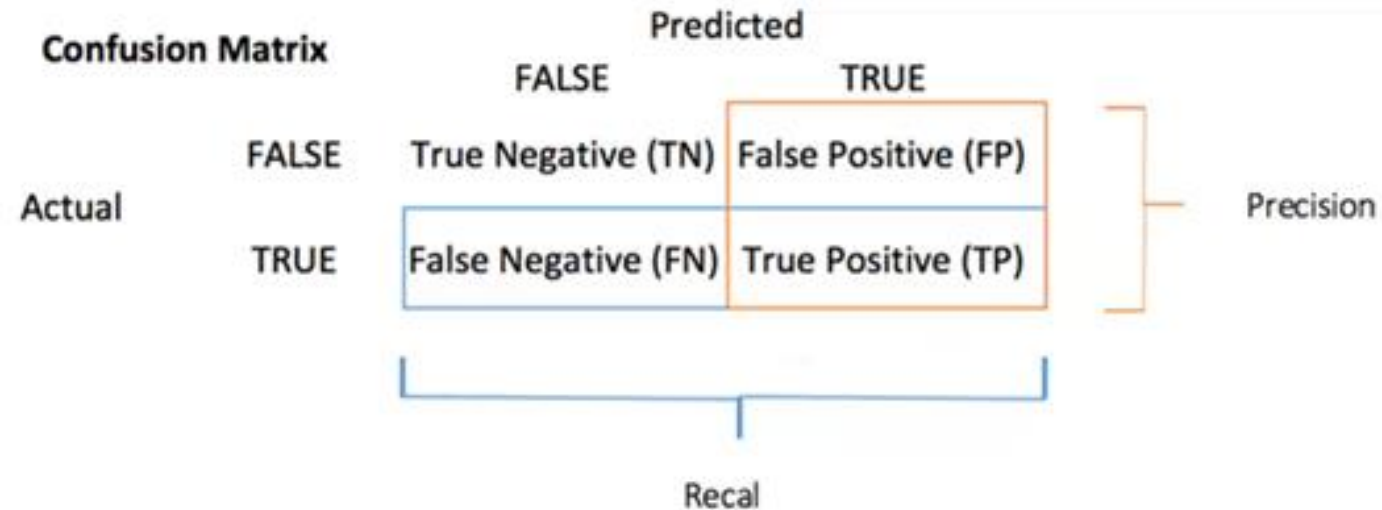
92%

Evaluation



Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Evaluation



- **Key Evaluation Metrics:**

- **Accuracy:**

- Overall correctness of predictions.

- **Precision:**

- Correct positive predictions out of all predicted positives.

- **Recall (Sensitivity):**

- Correct positive predictions out of all actual positives.

- **F1-Score:**


- Harmonic mean of precision and recall.

- **Key Considerations:**

- Use **Accuracy** for balanced datasets.
 - Focus on **Precision or Recall** for imbalanced datasets.
 - Prefer **F1-Score** for models requiring balance between precision and recall.

Table 1: Model Results with SMOTE								
	Models	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	MCC	ROC AUC
<u>20 Features</u>								
Cannabis	Extra Trees	0.9576	0.9576	0.9576	0.9576	0.9152	0.9152	0.9576
Alcohol	Extra Trees	0.9109	0.9109	0.9109	0.9109	0.8217	0.8217	0.9109
<u>Full Features</u>								
Cannabis	Extra Trees	0.9614	0.9616	0.9614	0.9614	0.9229	0.923	0.9614
Alcohol	Extra Trees	0.9236	0.924	0.9236	0.9235	0.8471	0.8475	0.9236
Model Results without SMOTE								
	Models	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	MCC	ROC AUC
<u>20 Features</u>								
Cannabis	Random Forest	0.8789	0.8703	0.8789	0.8739	0.4149	0.4173	0.6922
Alcohol	Random Forest	0.8153	0.8111	0.8153	0.8131	0.4119	0.4123	0.7012
<u>Full Features</u>								
Cannabis	ADA	0.8915	0.8757	0.8915	0.8768	0.3938	0.418	0.6576
Alcohol	Xgboost	0.8448	0.8316	0.8448	0.8328	0.4507	0.4632	0.698

How could we interpret the results and make policy suggestions?



Explainable AI (XAI)

Why Explainable AI?

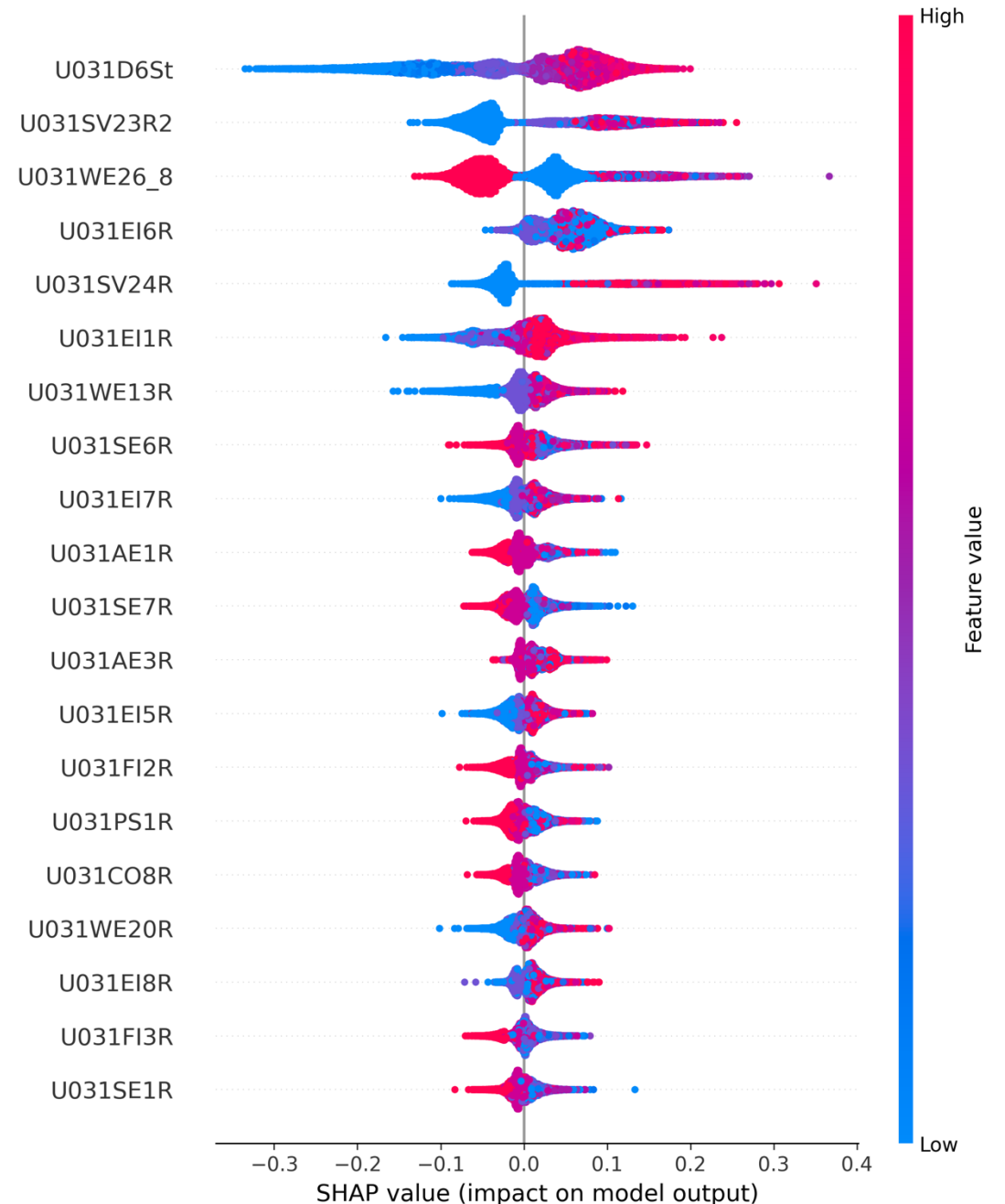
- Enhances **transparency** and **trust** in machine learning models.
- Identifies **feature importance** and explains predictions.

Key Tools for Model Interpretation:

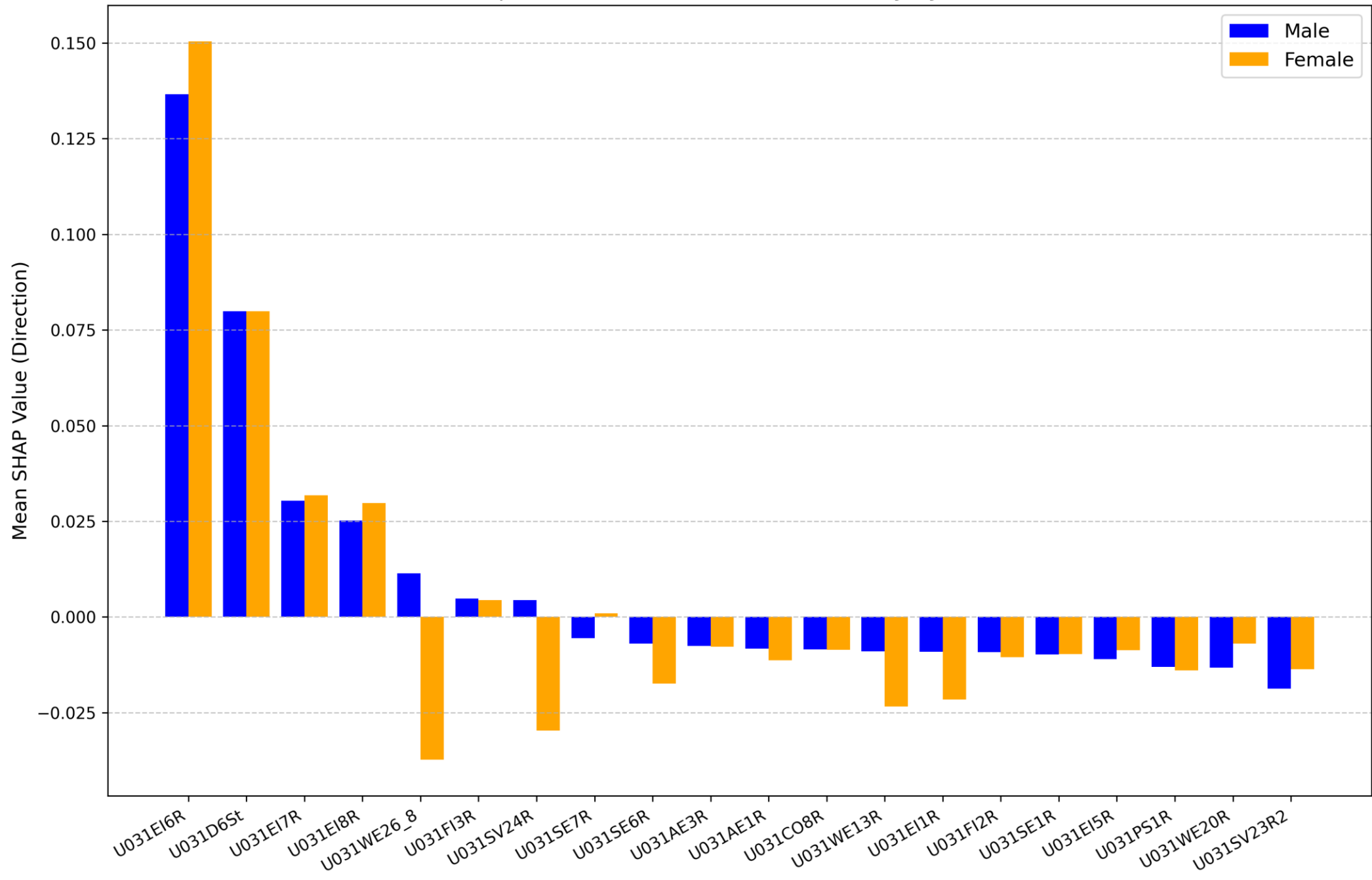
- **SHAP (SHapley Additive Explanations):**
 - Based on **game theory** to quantify each feature's contribution to predictions.
 - Provides **global** (model-level) and **local** (instance-level) explanations.
- **LIME (Local Interpretable Model-agnostic Explanations):**
 - Generates **local explanations** by approximating the model behavior near specific predictions.
 - Works with any type of model (**model-agnostic**).

Key Tip:

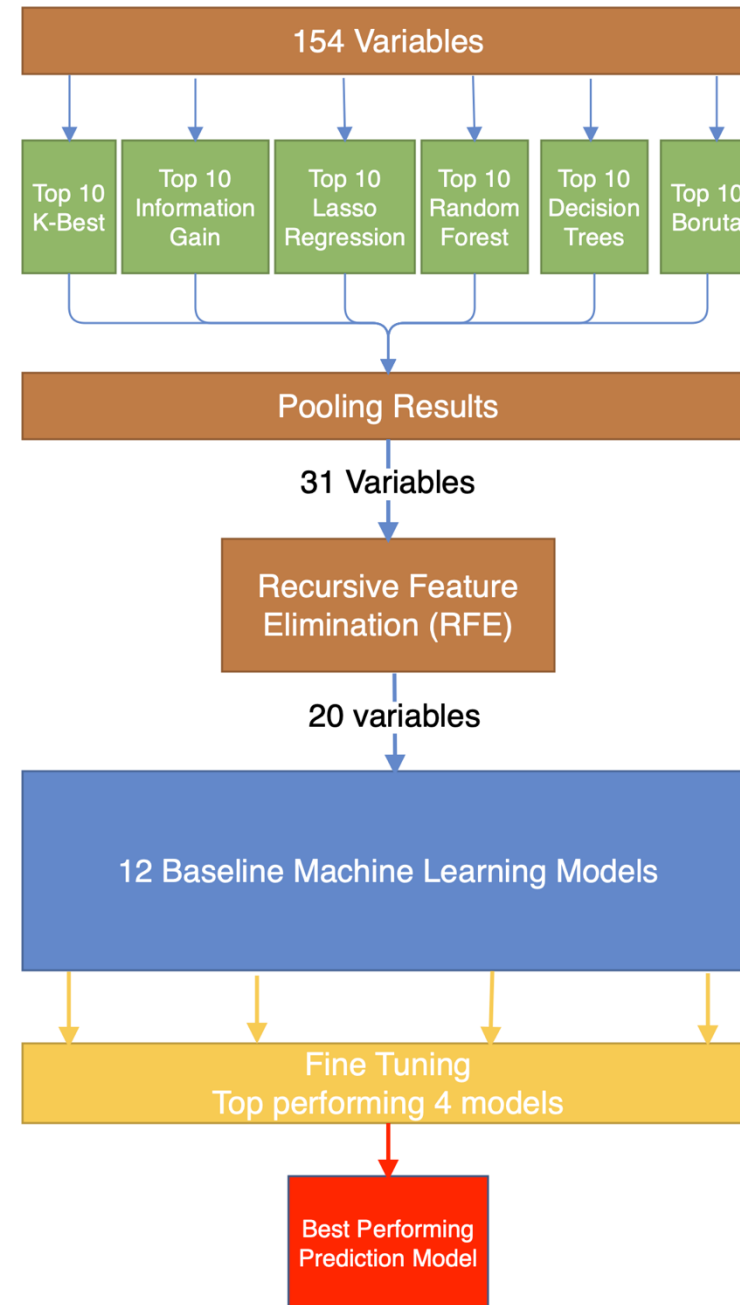
- Use SHAP for **detailed feature importance** and LIME for **quick interpretability** of specific predictions.



Comparison of SHAP Value Directionality by Gender



Overall Process





Unsupervised Learning

Cluster Analysis

Group individuals, behaviors, or entities based on similarities without predefined categories.

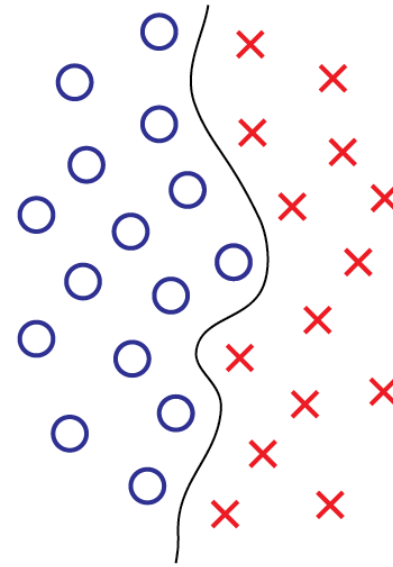
- **Sociology:** Identifying social groups based on shared characteristics, such as lifestyle preferences or political ideologies.
- **Economics:** Segmenting markets based on consumer purchasing behaviors.
- **Psychology:** Clustering individuals based on personality traits (e.g., Big Five traits).



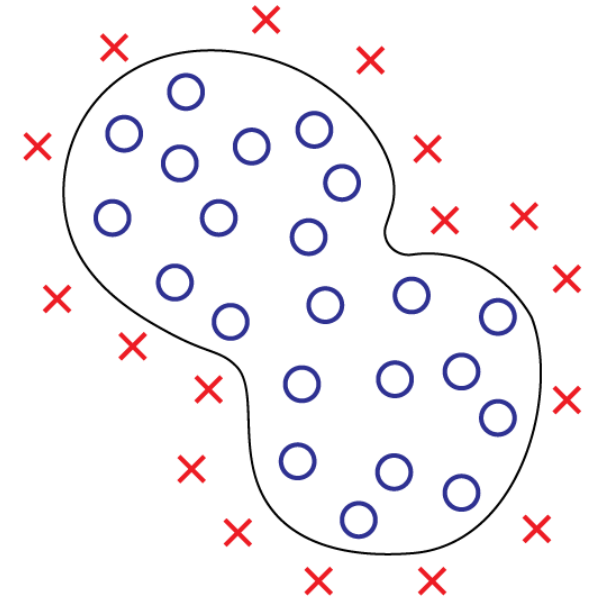
Anomaly Detection

Identify unusual or rare patterns in data.

- **Economics:** Detecting anomalies in economic trends or fraud detection in financial transactions.
- **Sociology:** Identifying deviant behaviors within social groups.
- **Health Studies:** Detecting unusual health patterns for epidemiological studies.



Classification

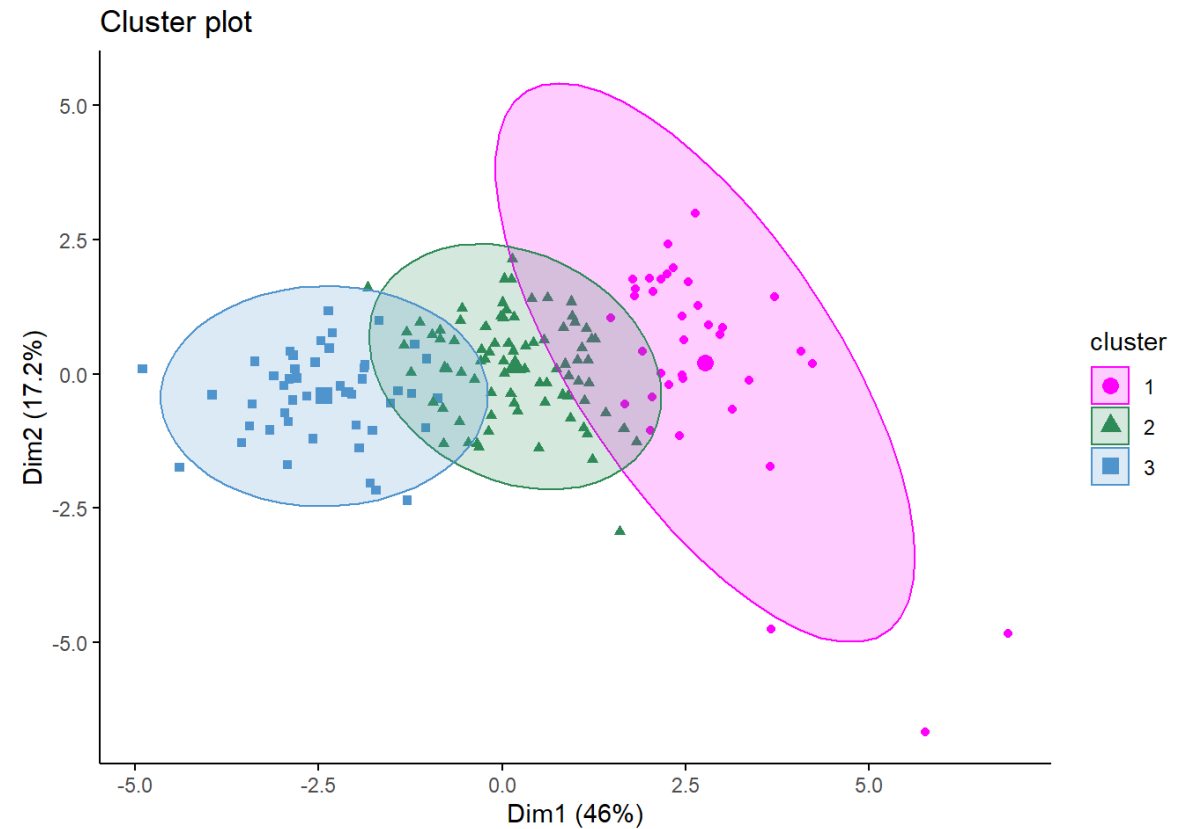


Anomaly Detection

Dimensionality Reduction (e.g., PCA)

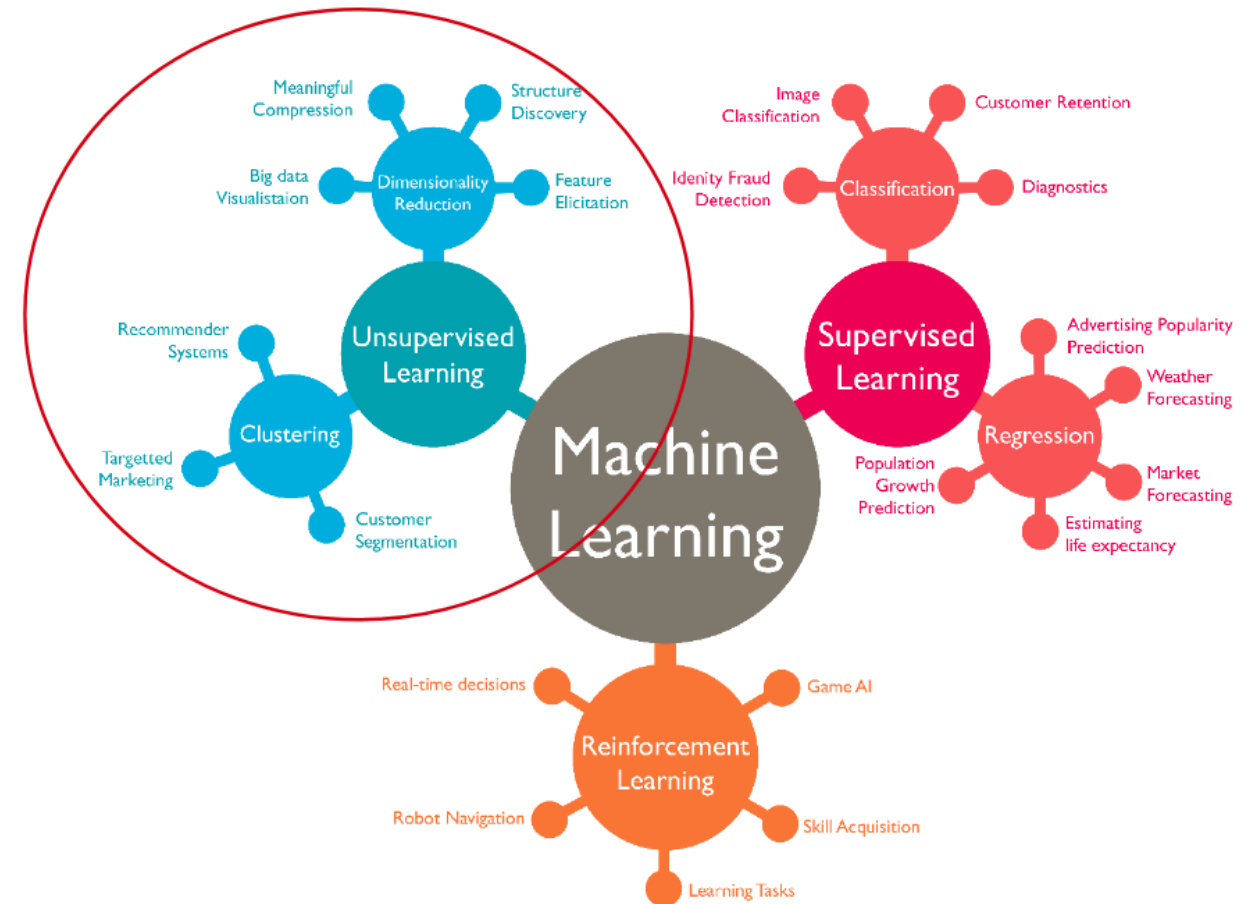
Reduce the complexity of large datasets while retaining important information, often for visualizing social patterns.

- **Survey Research:** Simplifying datasets with many variables to detect underlying factors (e.g., attitudes, preferences).
- **Cultural Analysis:** Identifying patterns in cultural artifacts or preferences based on high-dimensional data.
- **Political Science:** Mapping ideological spaces using data from voting records or survey responses.



Other Clustering methods

- **Topic Modelling:** Identify hidden topics or themes within textual data using algorithms
 - **Political Science:** Analyzing parliamentary debates to identify dominant topics discussed by politicians.
 - **Media Studies:** Extracting themes from news articles, tweets, or social media posts.
 - **Sociology:** Analyzing interviews or survey responses to detect underlying themes.
- **Network Analysis:** Discover hidden communities or structures in social networks.
 - **Sociology:** Analyzing social networks to identify cliques, influencers, or social roles.
 - **Communication Studies:** Mapping connections in online discussions or social media networks.
 - **Criminology:** Detecting organized crime networks or patterns in criminal activity.

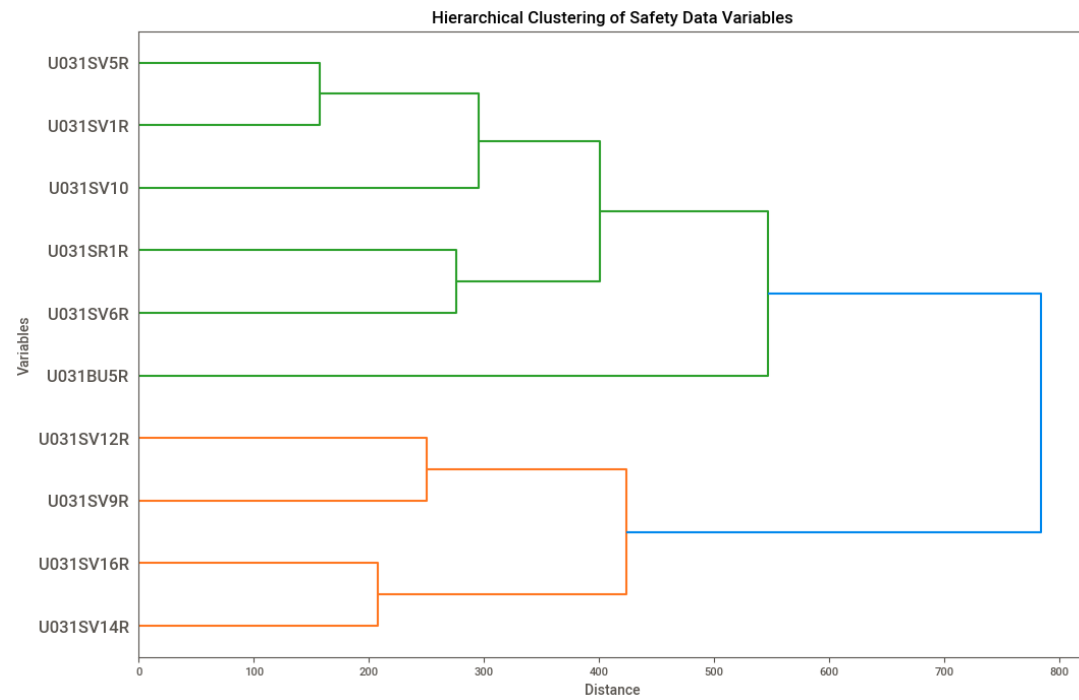


Dimensions of School Climate

- Goal:
 - First stage: reduce scale items from 30 to 9
 - Second stage: 9 to 3
- Clustering approaches
 - K-means
 - Multiple Correspondence Analysis (MCA)

Dimension	Item
Engagement	At this school, my teachers care about me
Engagement	At this school, my teachers listen to me when I have something to say
Engagement	My teachers believe that I can do well in school
Engagement	I enjoy learning at this school
Engagement	My teachers tell me when I do a good job
Engagement	At this school, I feel like I belong
Engagement	At this school, students and staff feel pride in this school
Engagement	At this school, students trust one another
Engagement	When I do something good at school, my parent(s) or guardian(s) usually hears
Engagement	The school provides instructional materials that reflect my culture, ethnicity, an
Environment	At this school, teachers can handle students who disrupt class
Environment	At this school, students listen to the teachers
Environment	At this school, it is easy for teachers at my school to control the students
Environment	Teachers at my school help students with their problems
Environment	At this school, there are clear rules about student behavior
Environment	At this school, everyone knows what the school rules are
Environment	At this school, students are rewarded for positive behavior
Environment	The school is usually clean and well-maintained
Environment	At this school, misbehaving students get away with it
Environment	There are often broken windows, doors, or desks in this school
Safety	Students carrying guns or knives
Safety	Harassment or bullying of students
Safety	Students' drug use (such as marijuana, LSD, cocaine, ecstasy)
Safety	Physical fighting between students
Safety	The students at my school use alcohol (such as beer, wine, liquor)
Safety	I feel safe at this school
Safety	I feel safe going to and from this school
Safety	Students at this school try to stop bullying
Safety	In the past 30 days, have you seen someone else being bullied?
Safety	The school has programs to deal with violence and conflict between students

Results



Questions & Comments

