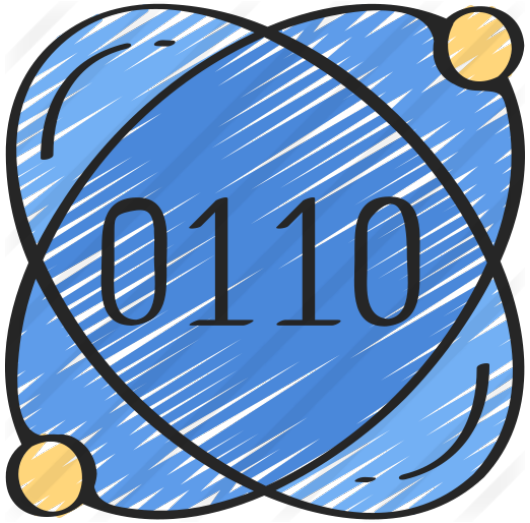
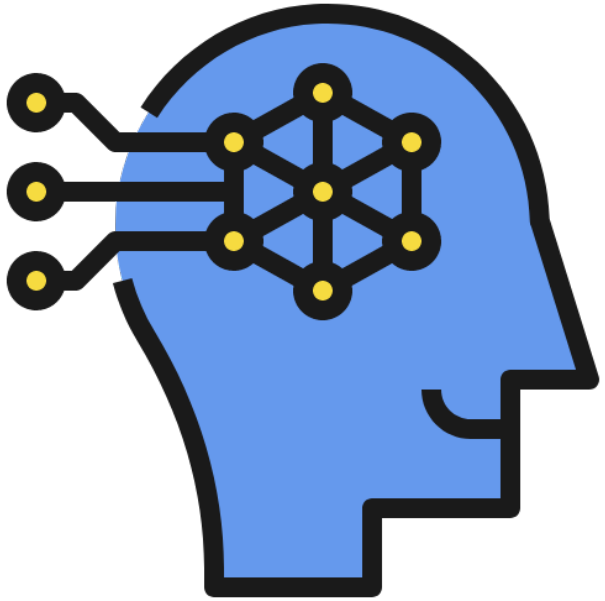


SAP Concur

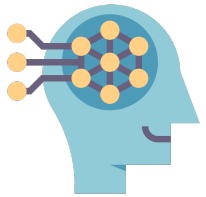
Prithvi Shetty - Data Scientist



Building an NLP machine learning model in 5 steps

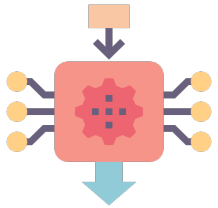
Overview

Background



Data Science and NLP

What interested me about NLP ?



Model building in 5 steps

Future scope

Background

MUMBAI UNIVERSITY

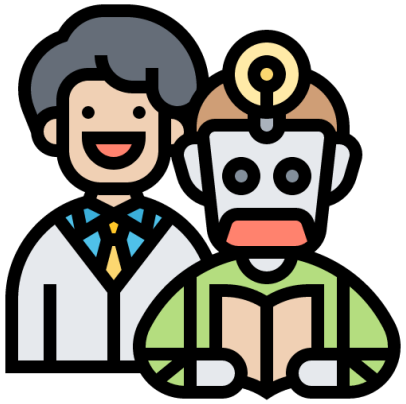
Undergrad Engineering

Masters at University of Washington

Data Scientist at SAP Concur

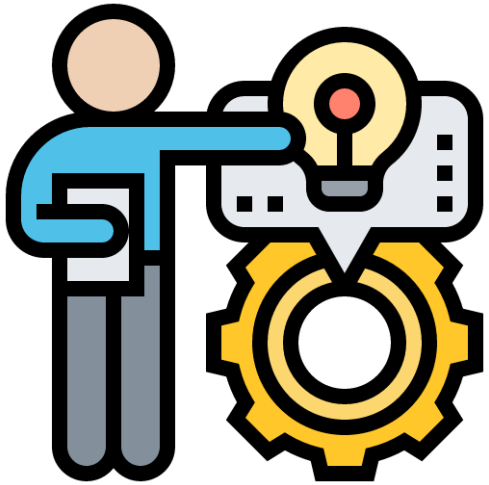
Data Science and Natural Language Processing

The 3 Types of Machine learning



Supervised

Unsupervised



Reinforcement

HEALTHCARE:
Patient Diagnosis

FINANCE:
Fraud Detection

MANUFACTURING:
Anomaly Detection

RETAIL:
Inventory Optimization

INSURANCE:
Client Risk Scoring

TRANSPORTATION:
Demand Forecasting

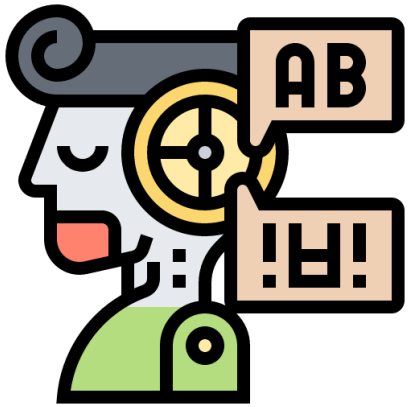
NETWORKS:
Intrusion Detection

E-COMMERCE:
Recommender Systems

MARKETING:
Customer Segmentation

ENERGY:
Demand Forecasting


Translation Application
Fake News Detection
Classifying Emails
Predicting Disease
Error Detection
IVR Application
Sentiment Analysis
Personal Voice Assistant



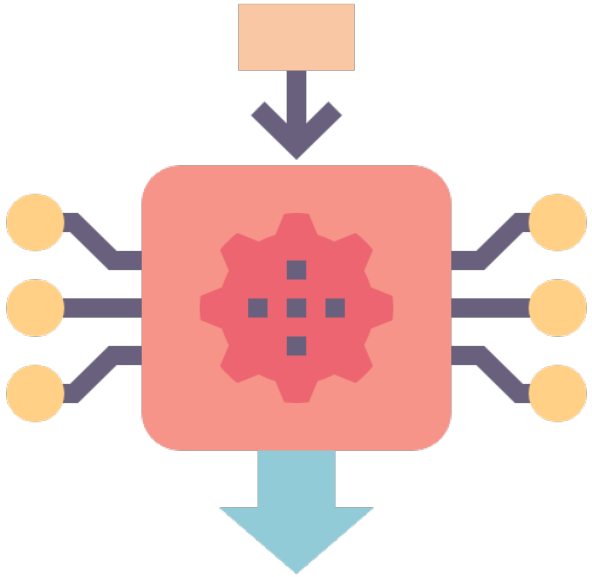
What interested me about NLP ?

Mapping the sentiment of people affected by Ebola in West Africa to mortality by region

Identifying early detection of Alzheimer's disease using NLP

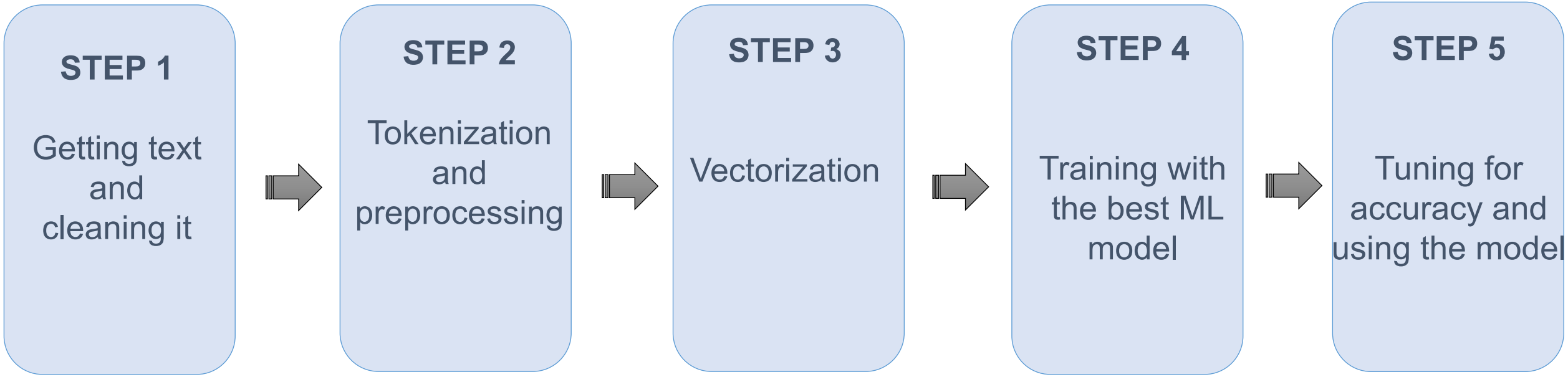
Text classification for Parkinson's disease

Data Science Internship at SAP

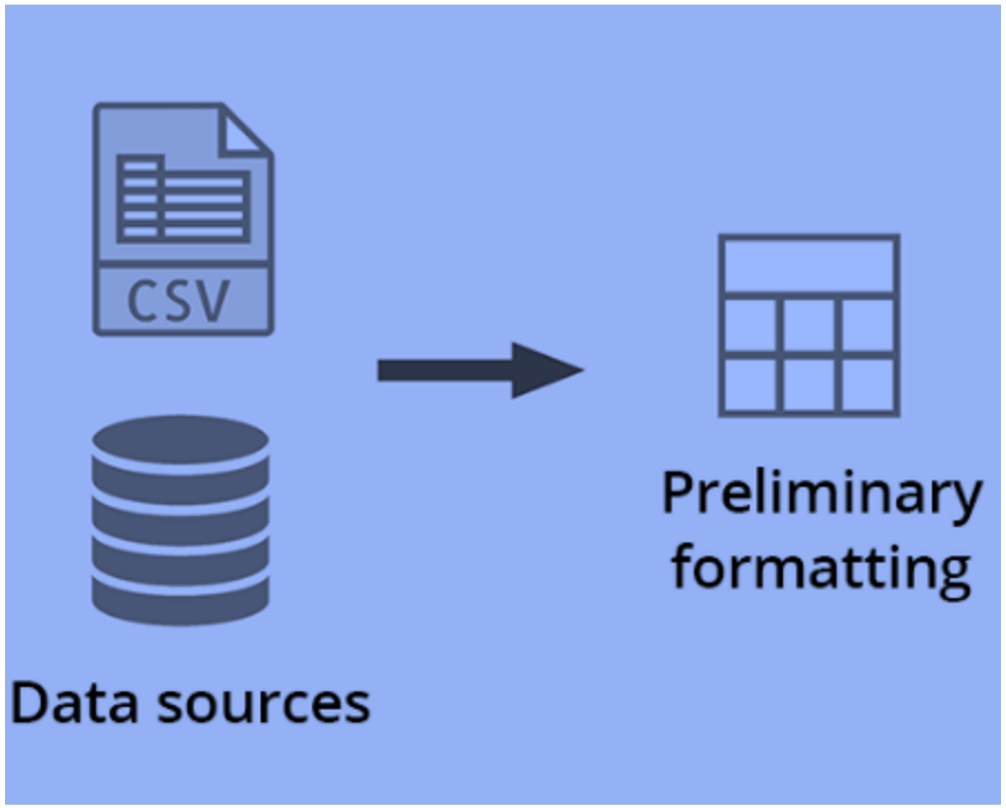


Model building in 5 steps

Process


STEP 1
Getting text and cleaning it
STEP 2
Tokenization and preprocessing
STEP 3
Vectorization
STEP 4
Training with the best ML model
STEP 5
Tuning for accuracy and using the model

Step 1 : Getting text data and cleaning it



Removing whitespaces

Manchester .

Manchester.

Lowercasing all words

MANCHESTER.

manchester.

Removing all non-alphanumeric characters

“Manchester” is located in UK.!.!.!>!

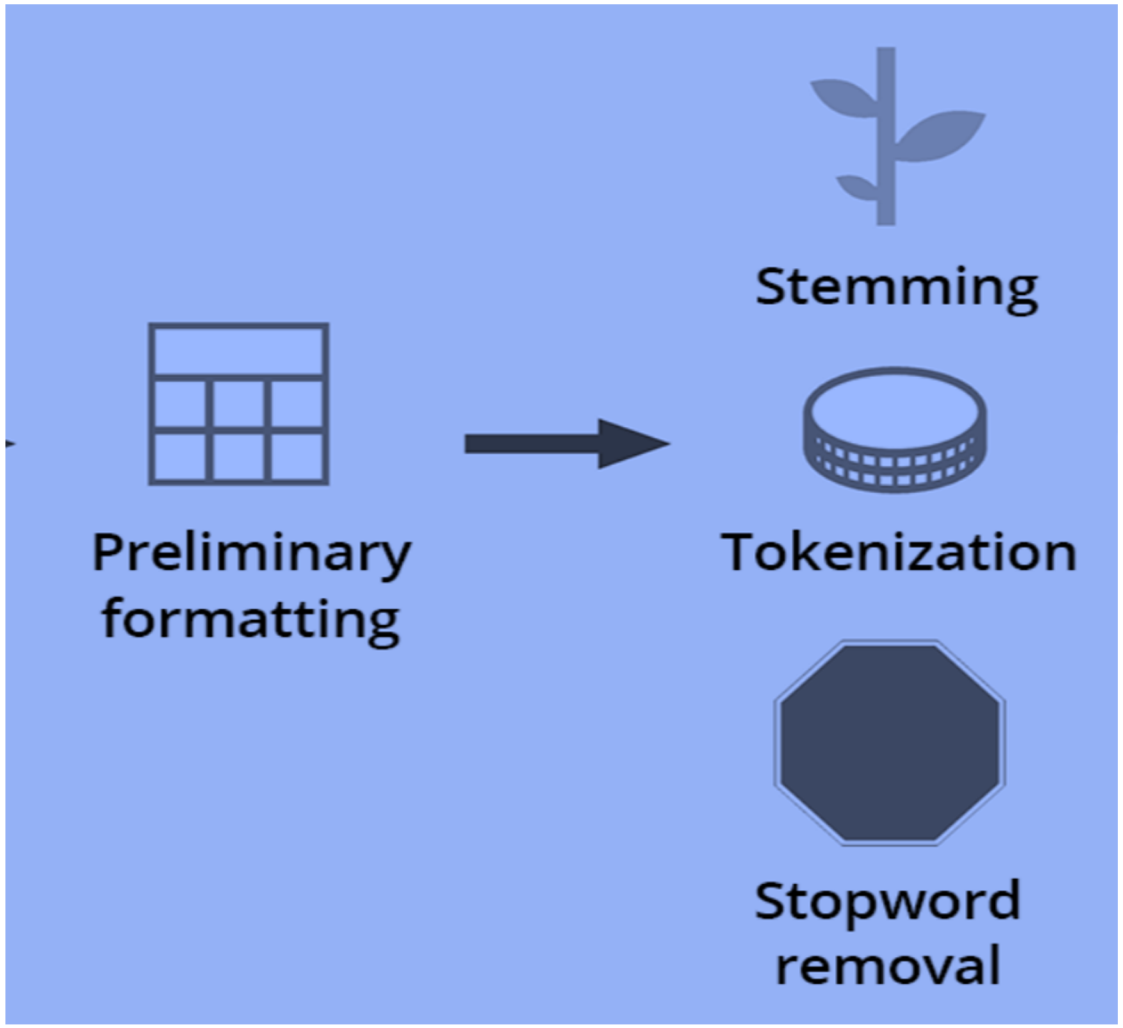
Manchester is located in UK

Before and after cleaning

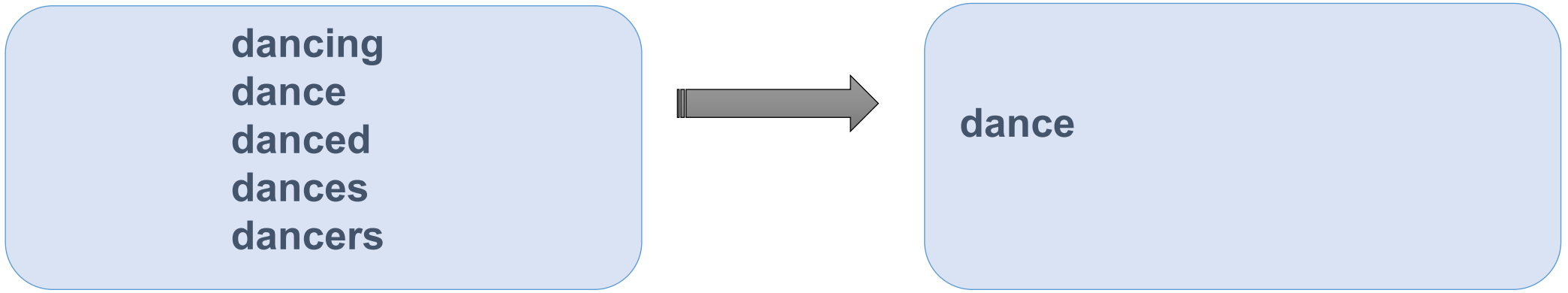
	class	text
0	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives around here though

	class	text
0	ham	go until jurong point crazy available only in bugis n great world la e buffet cine there got amore wat
1	ham	ok lar joking wif u oni
2	spam	free entry in a wkly comp to win fa cup final tkts st may text fa to to receive entry question std txt rate t c s apply over s
3	ham	u dun say so early hor u c already then say
4	ham	nah i don t think he goes to usf he lives around here though

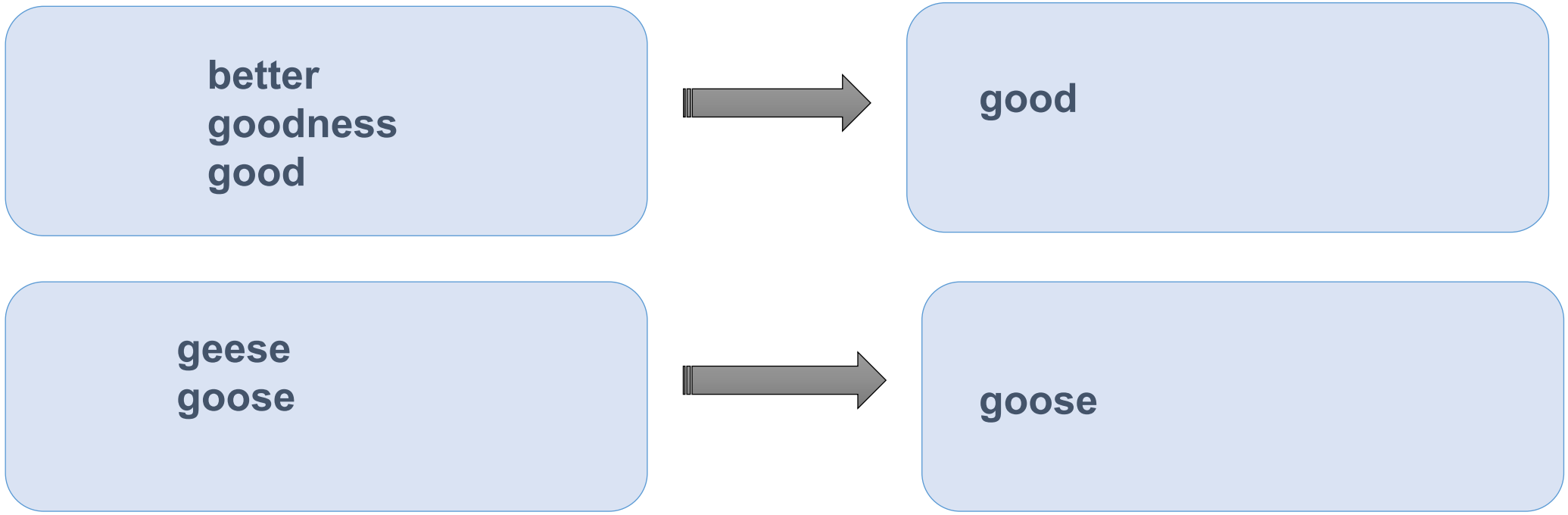
Step 2 : Tokenization and preprocessing



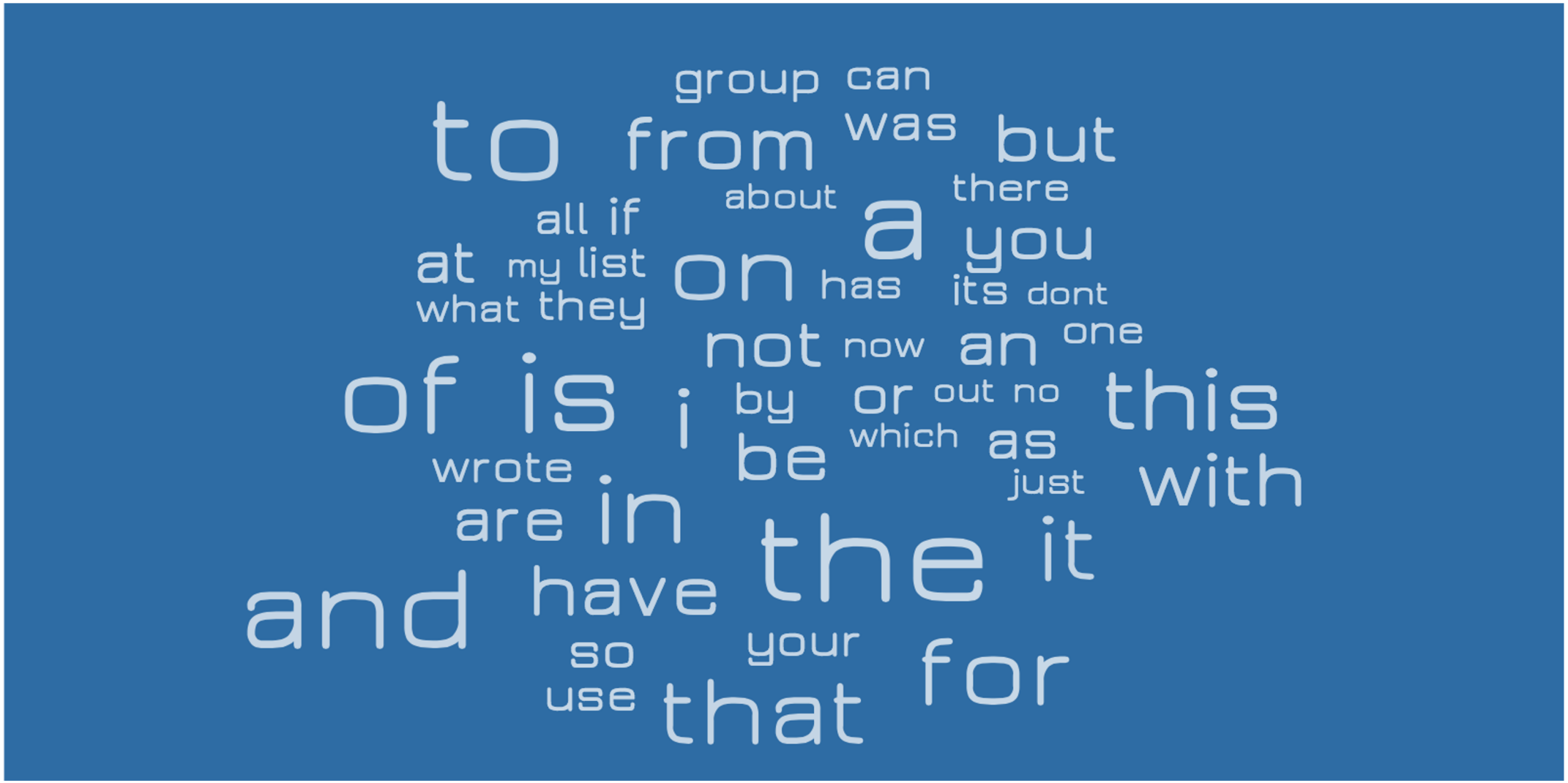
Stemming



Lemmatization



Removing stop words



Tokenization

“This is an example sentence”

[“This”,
“is”,
“an”,
“example”,
“sentence”]

Step 3 : Vectorization

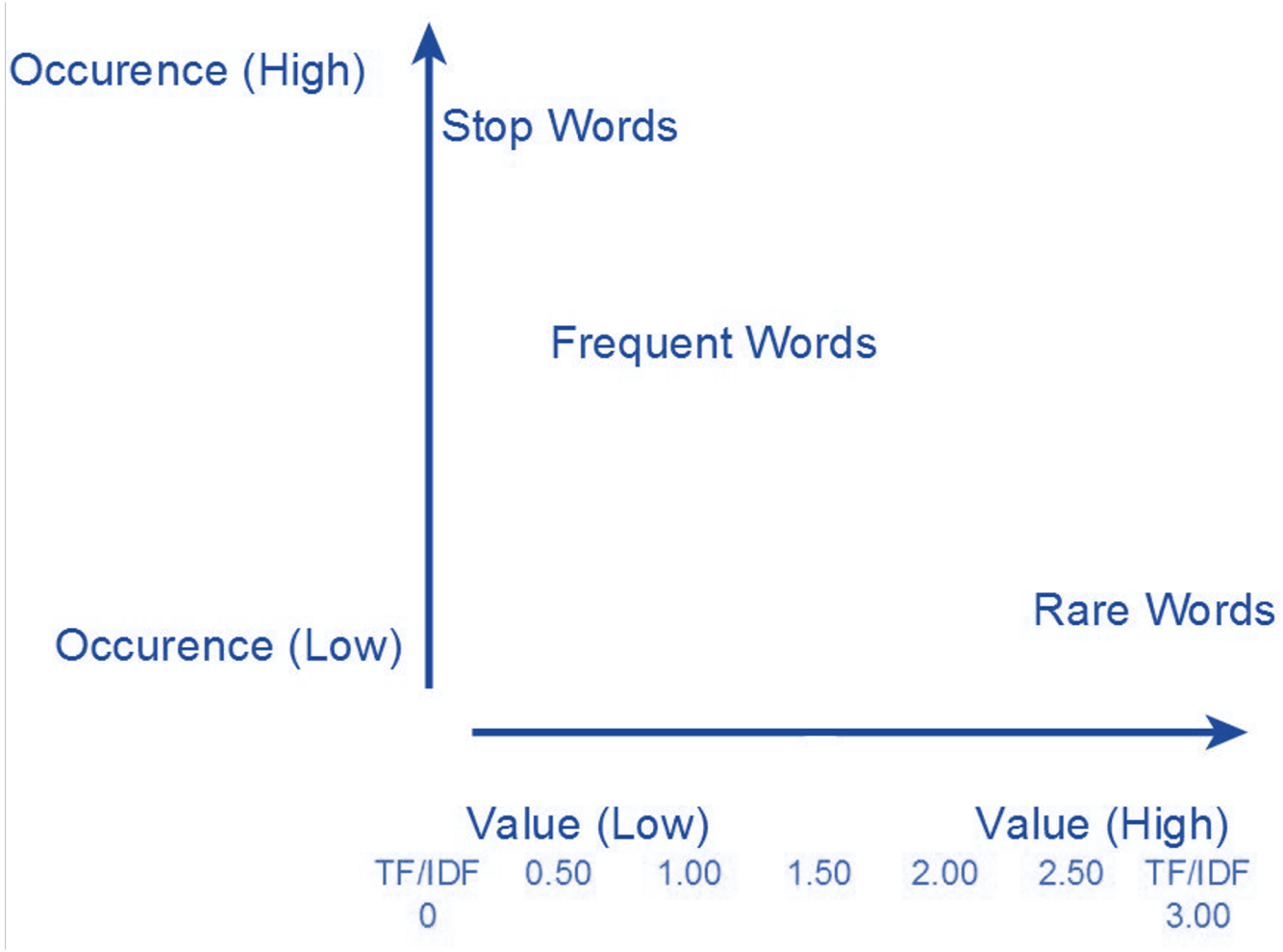
good movie
not a good movie
did not like

good	movie	not	a	did	like
1	1	0	0	0	0
1	1	1	1	0	0
0	0	1	0	1	1

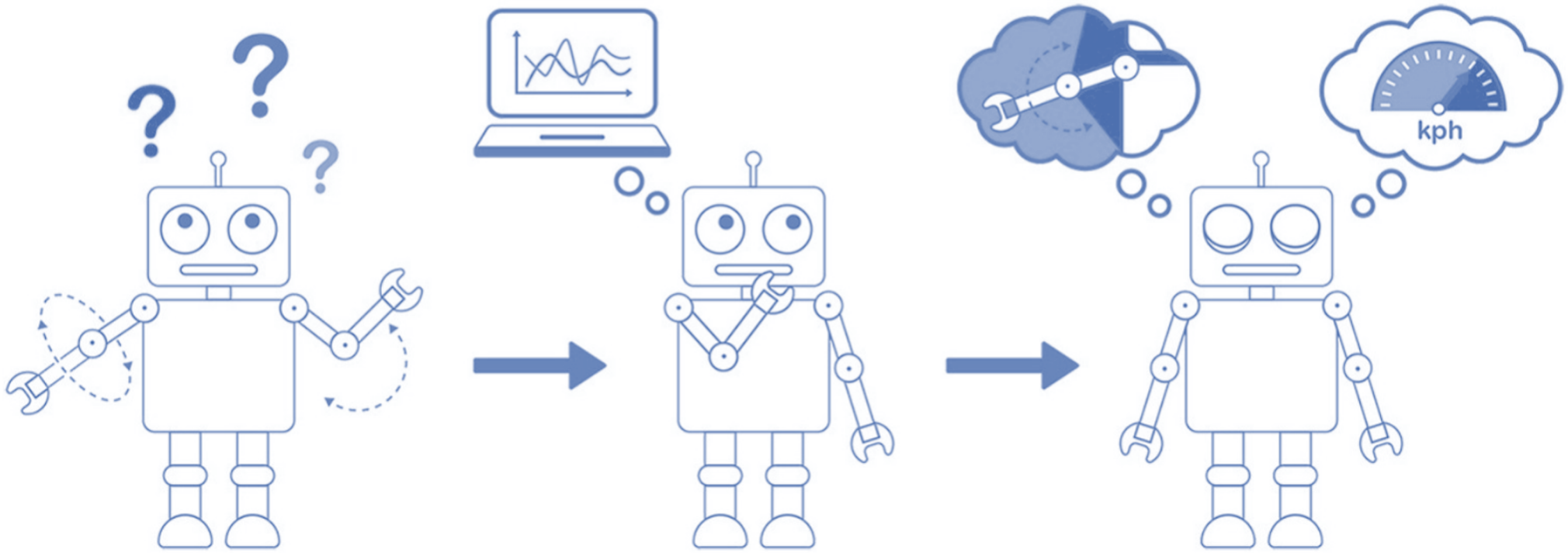
Bag-of-words approach (CountVectorizer)

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

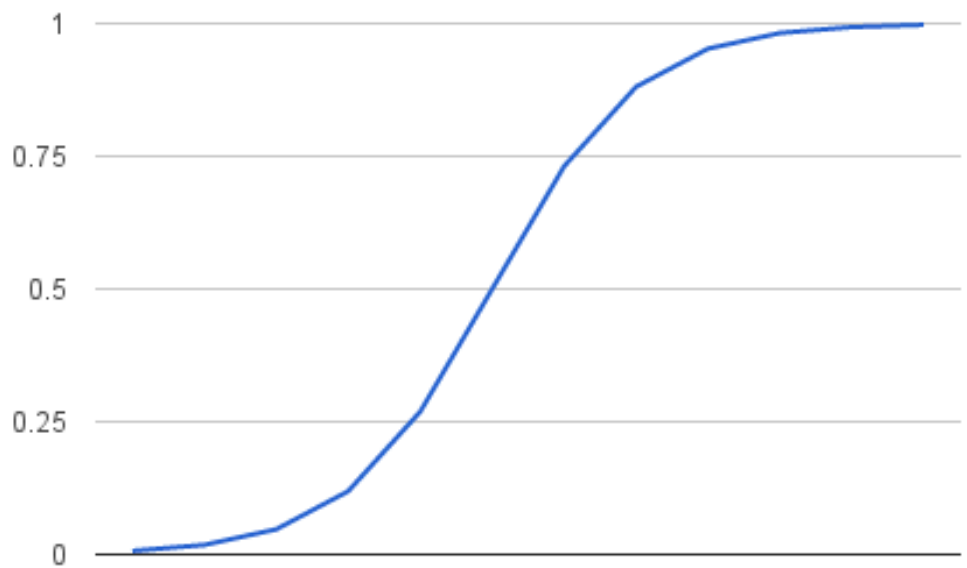
Tf-Idf (Term Frequency inverse document frequency)



Step 4 : Training with the best machine learning model



Logistic regression



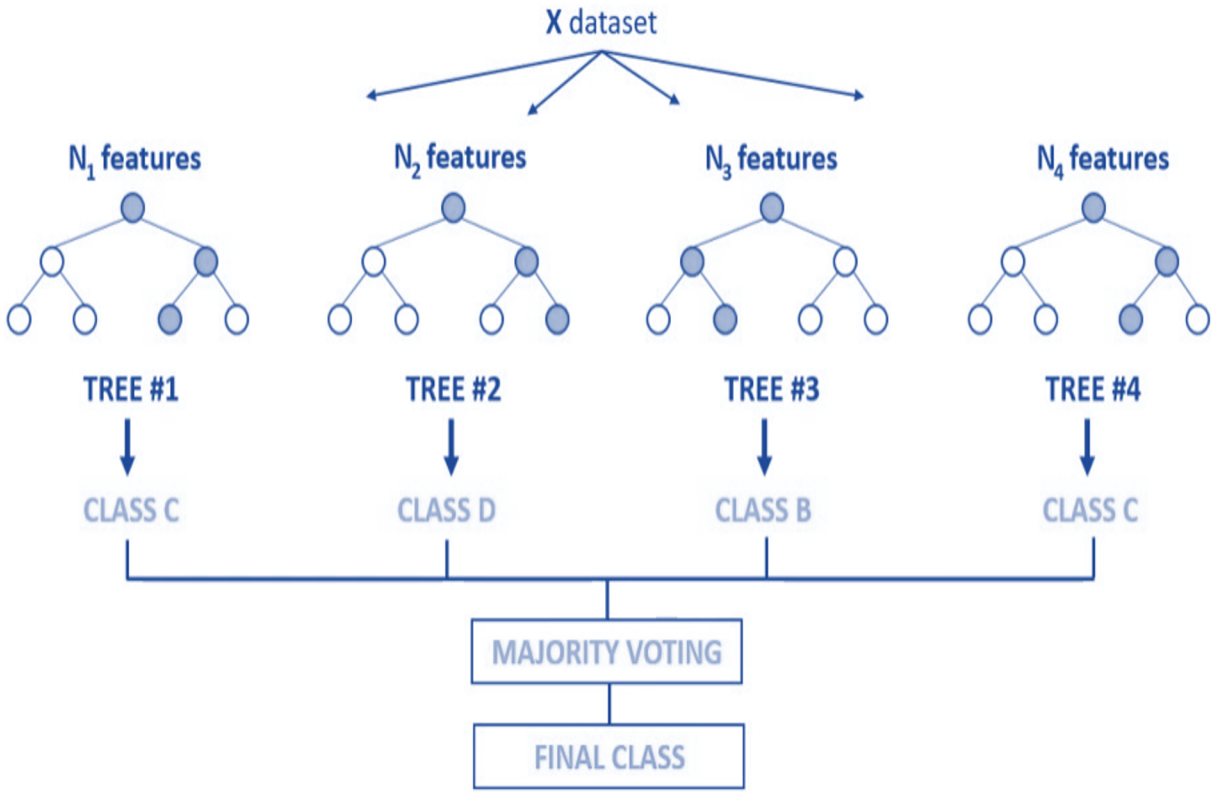
Naïve Bayes

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

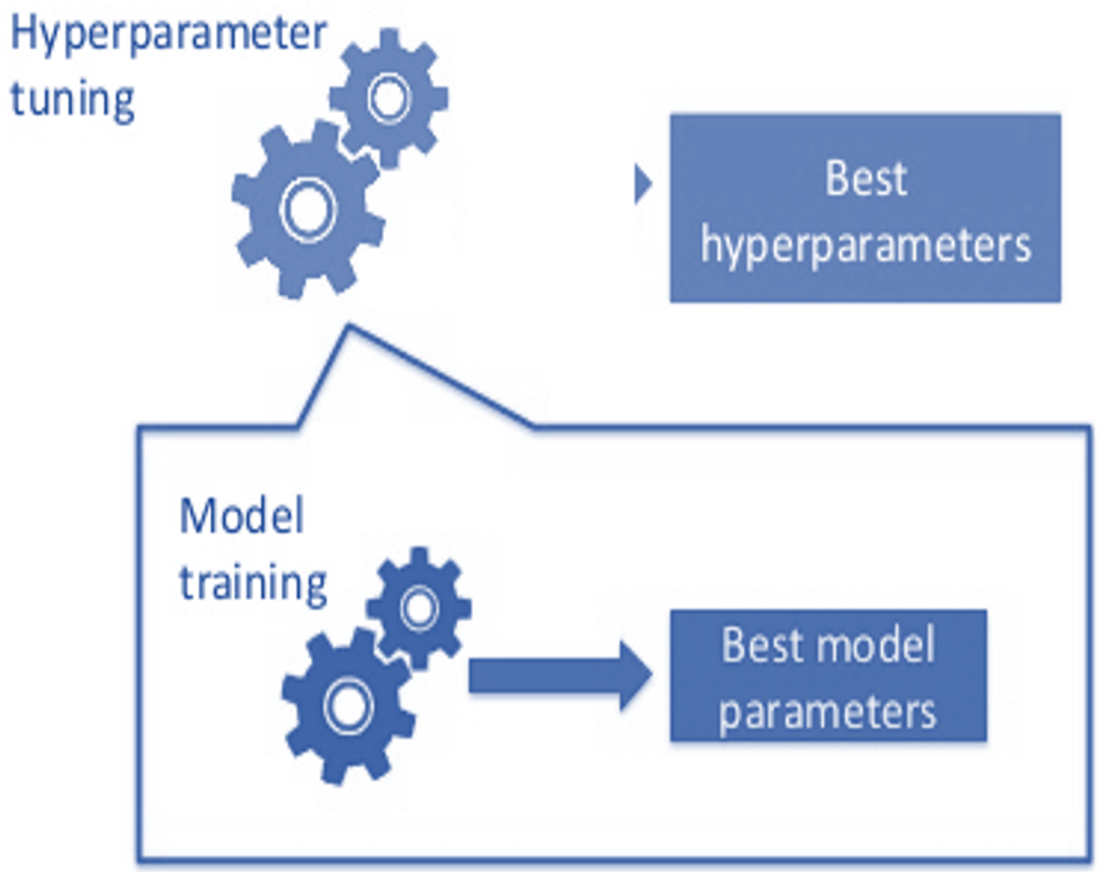
using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

Random Forest



Step 5: Tuning for accuracy and using the model



THANK YOU!

Please feel free to ask questions.

prithvi.shetty@sap.com