# Interpreting Code

With Natural Language Processing


## Eileen Cai

Data Scientist

# The problem

## Problem

Spread of child sexual abuse material online and other harmful content, such as, cyber bullying, drug sales and hate speech.

## Solution

Guardian of virtue: upholding digital ethics.

- Use machine learning to automatically detect harmful texts.

## Impact

Social impact on teen's mental health.
Companies benefit from:

- Sentiment Analysis
- Hate Speech Detection
- Content Moderation

# Preprocessing Procedures

| Train / Test Split | Re-sampling | Count Vectorization |
| --- | --- | --- |
| **Split data into train & test for modeling.** | **To counter for class imbalance.** | **Transform text into numbers.** |

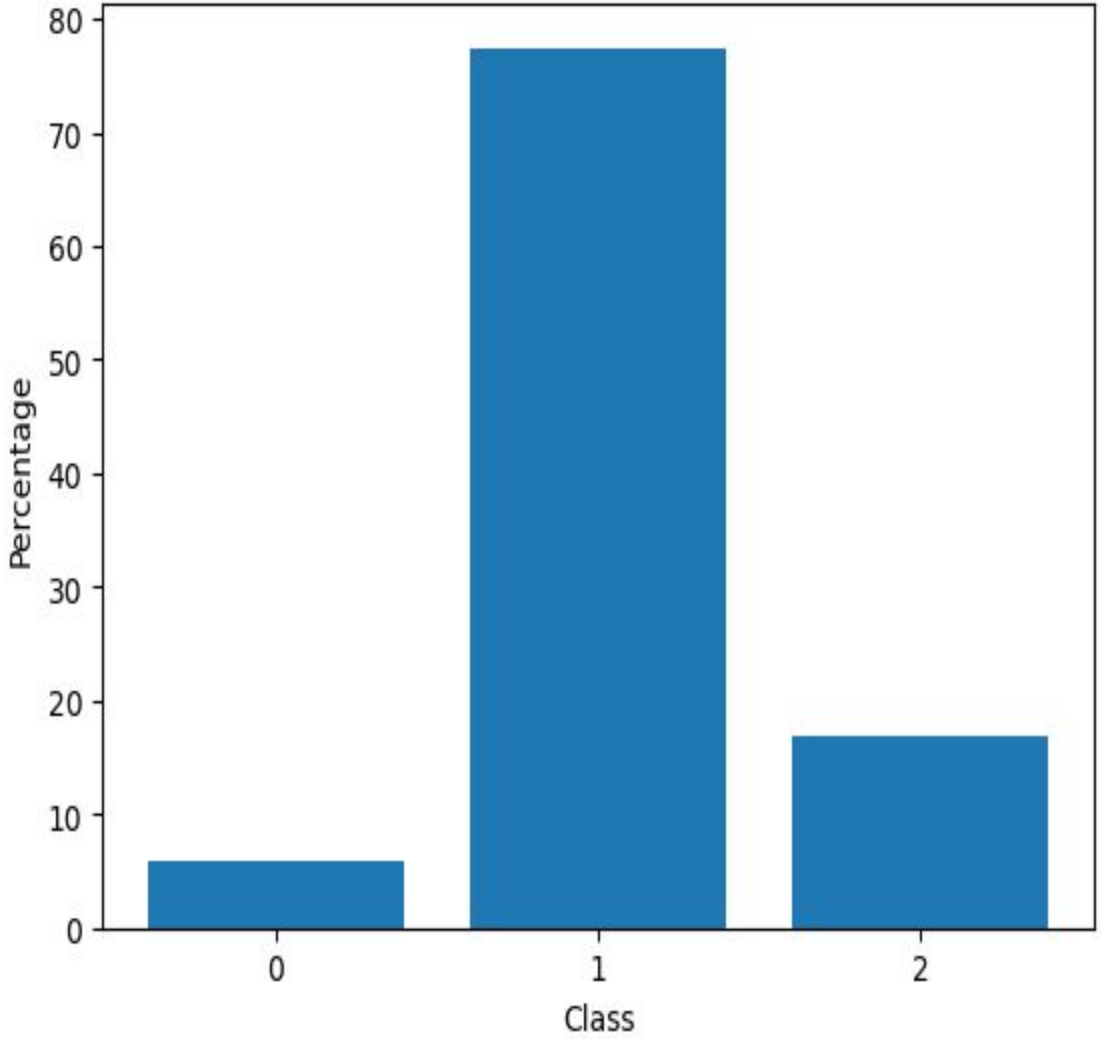
Target: class
Feature: tweet

Data source: Huggingface hate_speech_offensive

# EDA

Class Imbalance

- class 0 (6%) - hate speech
- class 1 (77%) - offensive languag
- class 2 (17%) - neither

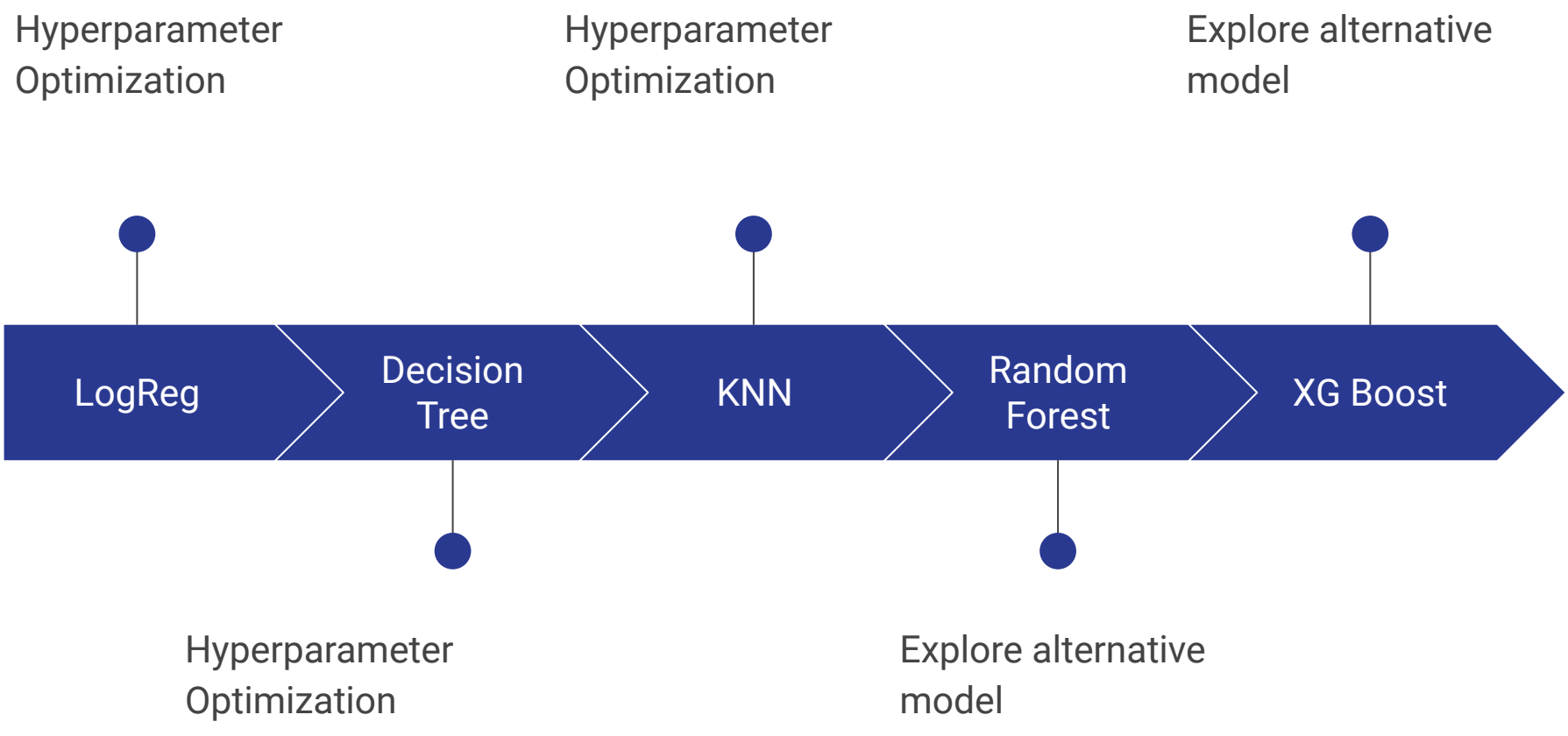


# Baseline models

Logistic Regression
Decision Tree Classifier
KNN Classifier

# Evaluation Metrics

Precision
Recall
F1-score


Hyperparameter Optimization
Hyperparameter Optimization
Explore alternative model
LogReg
Decision Tree
KNN
Random Forest
XG Boost
Hyperparameter Optimization
Explore alternative model

# Questions