Phase 3 Project

Author: Tommy Phung

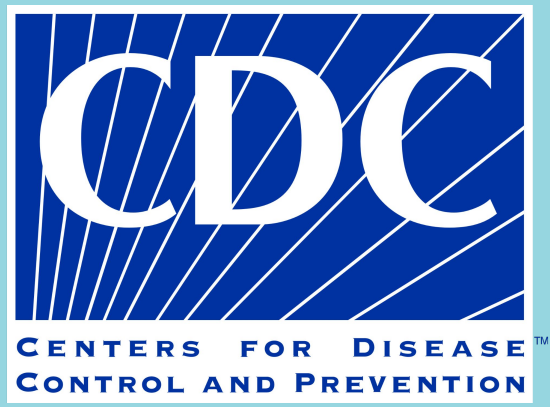
Overview

Client: Providence and Medical Centers

Source: CDC

Target Variable: Seasonal Vaccines

Goal: Create a model to predict whether a patient will take the seasonal vaccine.



Business Problem

Vaccines are wasted due to:

- ➔ **Expire date**
- ➔ **Supply Chain Issue**

Real Life Example:

- ➢ **1.1 billion** Covid Vaccines were estimated to be **wasted** due to expired vaccines and supply chain issues.

Data Understanding

Number of observation: 26,000 participants

➔ **36 → 105** Columns (after preprocessing)

➔ **24** numeric columns

➔ **12** object columns

Target Variable: Seasonal Flu Vaccine

Method

Preprocessing Steps

- Split Dataset to **Training** and **Testing** Sets
- Prepare Datasets
  - (Missing values, Scalar, Dummy Variables)
- **Fit Model** with Training Dataset
- Make **Predictions**

Models

- **Simple Baseline Model**

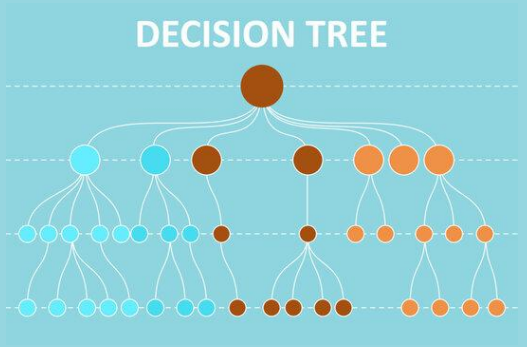
  - Decision Tree

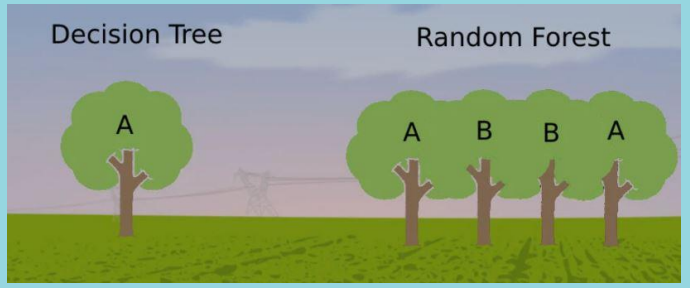
- **Complex Model**

  - Random Forest

- **Tuned Model**

  - Random Forest with Tuned Hyperparameters





Simple Decision Tree

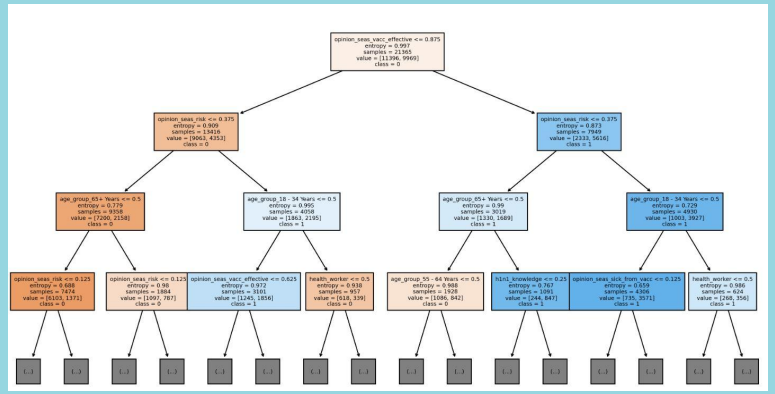
Decision Tree Classification:

- Split dataset based on features in order to reduce entropy.

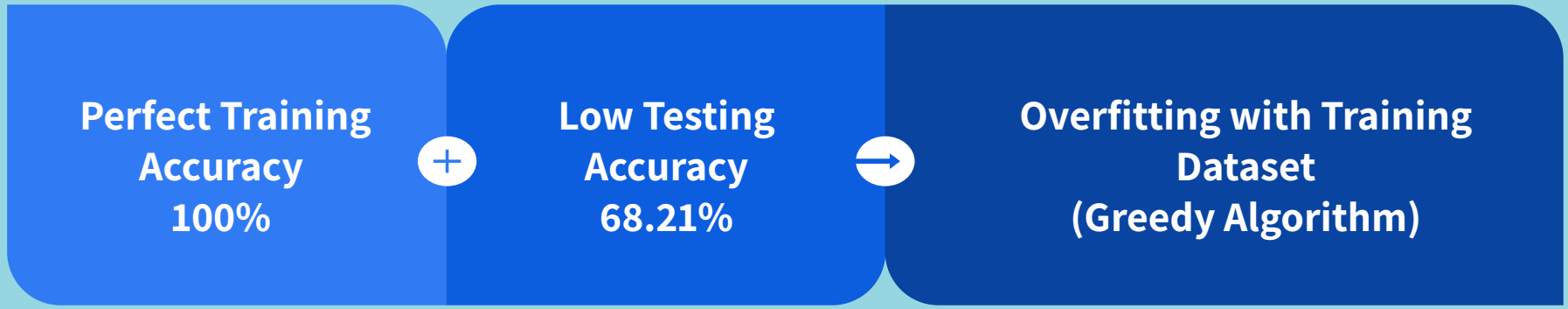
Parameters: Default, criterion = entropy

Training Accuracy: **100%**

Testing Accuracy: **68.21%**



Decision Tree - Analysis



Overfitting results in larger difference in accuracy

Decision Tree → Random Forest

Decision Tree

Pro:

➔ Easy to Interpret

Con:

➔ Prone to overfitting

Random Forest

Pro:

➔ Resilient to overfitting

➔ Resistance to noise

Con:

➔ Long Computational Time

Random Forests

Random Forest Classification:
Create multiple Decision Trees with different set of features.

Parameters:
Max Depth = 10, Max Features = None, criterion = entropy

Training Accuracy: **83.45%**
Testing Accuracy: **77.89%**

Random Forest with Tuned Hyperparameters

Tuning with Grid Search

- Compares models with different parameters
- Over **7,000** combinations

Training Accuracy: **92.59%**
Testing Accuracy: **78.36%**

	Default	Tuned
n_estimators	100	**300**
criterion	**Entropy**	**Entropy**
max_depth	**None**	**15**
min_samples_splits	2	**1**
min_sample_leaf	1	**1**
max_features	Sqrt (7)	**15**

Model Comparison

Model	Training Accuracy	Testing Accuracy	Recall Score
Decision Tree	**100%**	**68.21%**	**65.33%**
Random Forest	**83.45%**	**77.89%**	**75.55%**
Random Forest with hypertuning	**92.59%**	**78.36%**	**75.55%**

Real Life Application (Example)

Data Taken from Testing Dataset

Number of Patients	Predicted Taken	Patients Actually Taken	Number of Vaccines + 5% *	Vaccines Wasted
5342	2416	2466	2416 + 121 = 2537	71

* Supply should be slightly more in case of faulty or new patients
Subject to change based on location. (Major Cities, Small Towns)

Conclusion

- Random Forest Performs **Better** Than Decision Trees
- **92.59% Training Accuracy**
- **78.36% Testing Accuracy**

- Use Model to obtain rough estimate number of vaccines
- **71** wasted vaccines vs **2,876** *

*** Assuming everyone needs a vaccines**

Next Steps

1. **More Tuning**
2. **Different Classification Models**
    a. KNN
    b. XGBoost
3. **Better Survey Questions**
    a. Demographic
    b. Religion

Questions?

Number of Patients	Predicted Taken	Patients Actually Taken	Number of Vaccines + 5% *	Vaccines Wasted
5342	2416	2466	2537	71

Model	Training Accuracy	Testing Accuracy	Recall Score
Decision Tree	**100%**	**68.21%**	**65.33%**
Random Forest	**83.45%**	**77.89%**	**75.55%**
Random Forest with hypertuning	**92.59%**	**78.36%**	**75.55%**