Phase 3 Project

Author: Tommy Phung

Overview

Client: Providence and Medical Centers

Source: CDC

Target Variable: **Seasonal Vaccines**

Goal: **Minimize** vaccine wastage when ordering vaccines.

Business Problem

Vaccines are wasted due to:

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

Real Life Example:

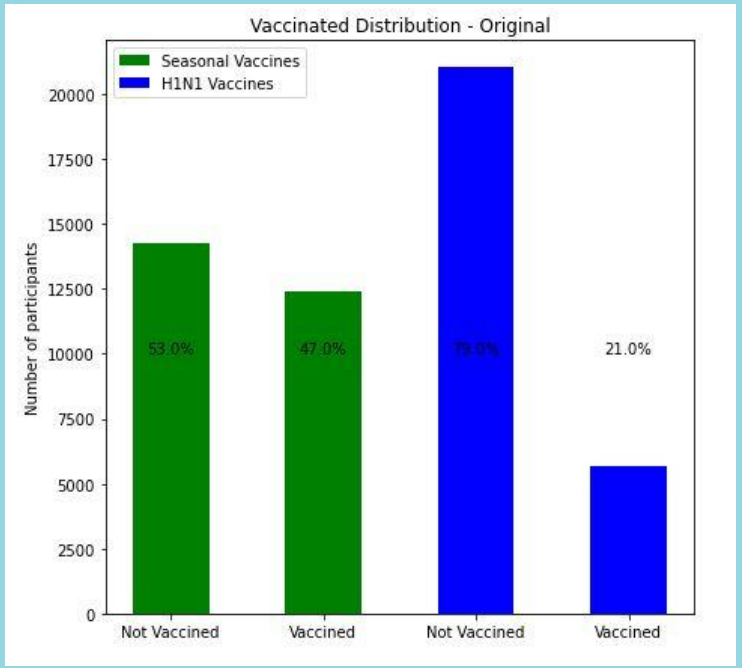
- ➢ **1.1 billion** Covid Vaccines were estimated to be **wasted** due to expired vaccines and supply chain issues.

Data Understanding

Number of observation:
26,000 participants

- ➔ **36** different survey question
- ➔ Roughly **50%** taken seasonal vaccine

Target Variable: <u>Seasonal Flu Vaccine</u>



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**

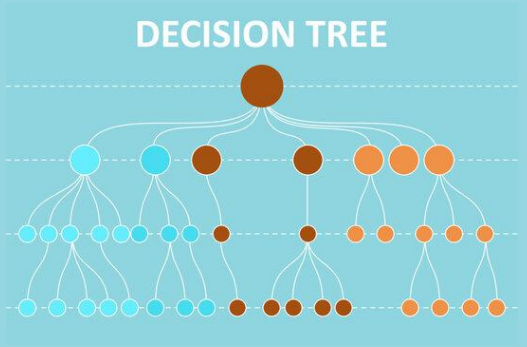
  - Decision Tree

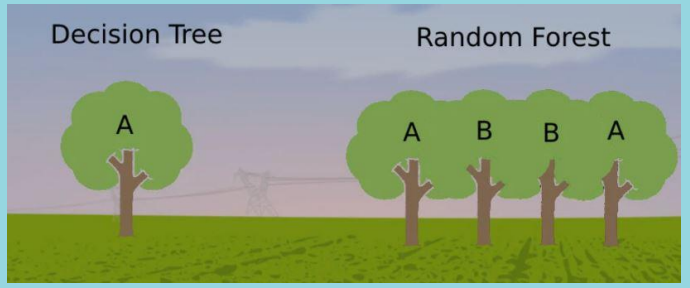
- **Complex Model**

  - Random Forest

- **Tuned Model**

  - Random Forest with Tuned Hyperparameters





Simple Decision Tree

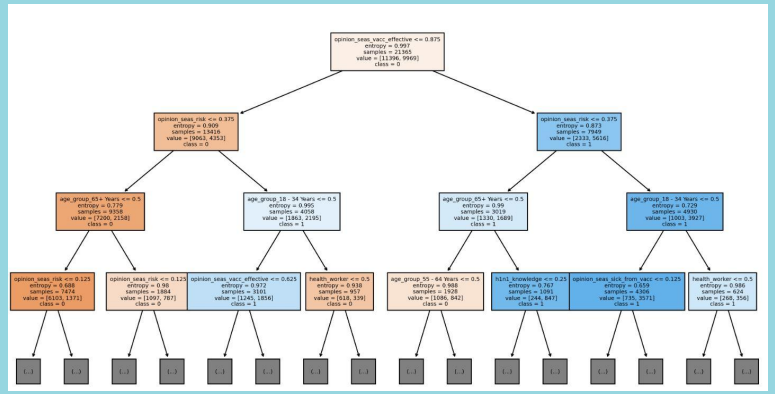
Decision Tree Classification:

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

Parameters: Default, criterion = entropy

Training Accuracy: **100%**

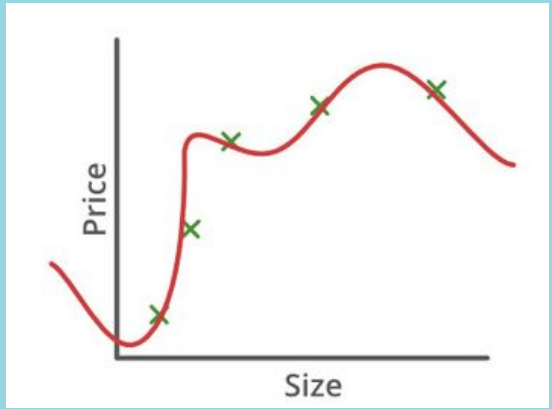
Testing Accuracy: **68.21%**



Decision Tree - Analysis

Perfect Training Accuracy 100% + **Low Testing Accuracy 68.21%** → **Overfitting with Training Dataset (Greedy Algorithm)**

Overfitting - A model trained to perfectly predict the training dataset



Decision Tree → Random Forest

Decision Tree	Random Forest
+ **Interpretable**	+ Resilient to **overfitting**
- **Prone** to overfitting	+ Resistance to **noise**
	- **Long** Computational Time

Random Forests

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

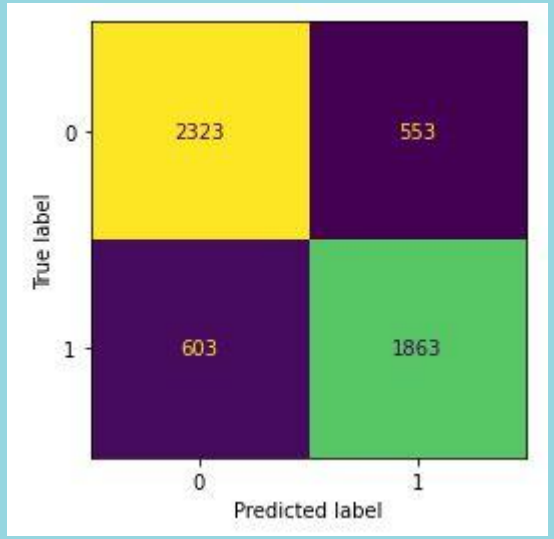
Parameters:
Max Depth = 10, Max Features = None, criterion = entropy

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%

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)

Recommendation

Most Important Features:
Opinion_Seasonal_Risk, Opinion_Seasonal Vaccine_Effective, Doctor_Recommendation_Seasonal

Question Concerning Vaccines → **More Likely** to take vaccine

Recommendation:

- Only include question regarding **current** vaccine focus
- Occupation have **little to no** influence in prediction

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

Thank You…Questions?

Email: phungtommy109@gmail.com
GitHub: https://github.com/Tommyphung1/phase_3_project