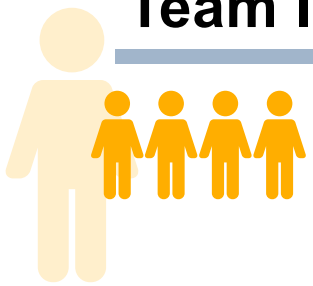
# What features of potential candidates should be assessed for Covid-19 vaccination ?

Jia HUANG, Siyu CHEN, Xia HU, Shuhui CHEN, Gefei LI

2022.08.09

# CONTENTS

# Team Infrastructure



| | |
|---|---|
| Siyu CHEN | Coordination, Logic, Data Processing, Programming |
| Xia HU | Data processing, Programming |
| Jia HUANG | Database, Models, References |
| Shuhui CHEN | Conclusion, Evaluation, Data Analysis |
| Gefei LI | Data Processing, Programming |

/01

# Introduction

# Problem Statement

01 Severe COVID-19

02 Vaccine Development

03 Negative Effects

Death

**Motivation**

**Learn whether the life of a candidate is dangerous once taken the COVID-19 vaccine**

# Literature Review


01

## Zhu et al. (2021)

- Predict whether a patient would develop severe symptoms of COVID-19 later.

02

## Iwendi et al. (2020)

- Predict the severity of the case and the possible outcome, recovery, or death.

03

## Shahid et al. (2021)

- A high demand for ML-aided diagnosis systems
- Predict the spread of COVID-19

/02

# EDA

# Data Description

| Output Features | | Input Features | | | |
|---|---|---|---|---|---|
| **Vaccine Symptoms** | **Description** | **Population Features** | **Description** | **Type of Vaccine** | **Description** |
| DIED* | Death count | AGE_YRS | Age of samples | VAX_MANU | Vaccine manufacturer (MODERNA/PFIZER) |
| L_THREAT | Whether people feel the life threatening | SEX | Sex (female/male) of samples | VAX_DOSE_SERIES | How much dosage of vaccine people accept |
| RECOVD | Whether people suffer from side effect recovered | | | | |

# Binary Description

## Input Features

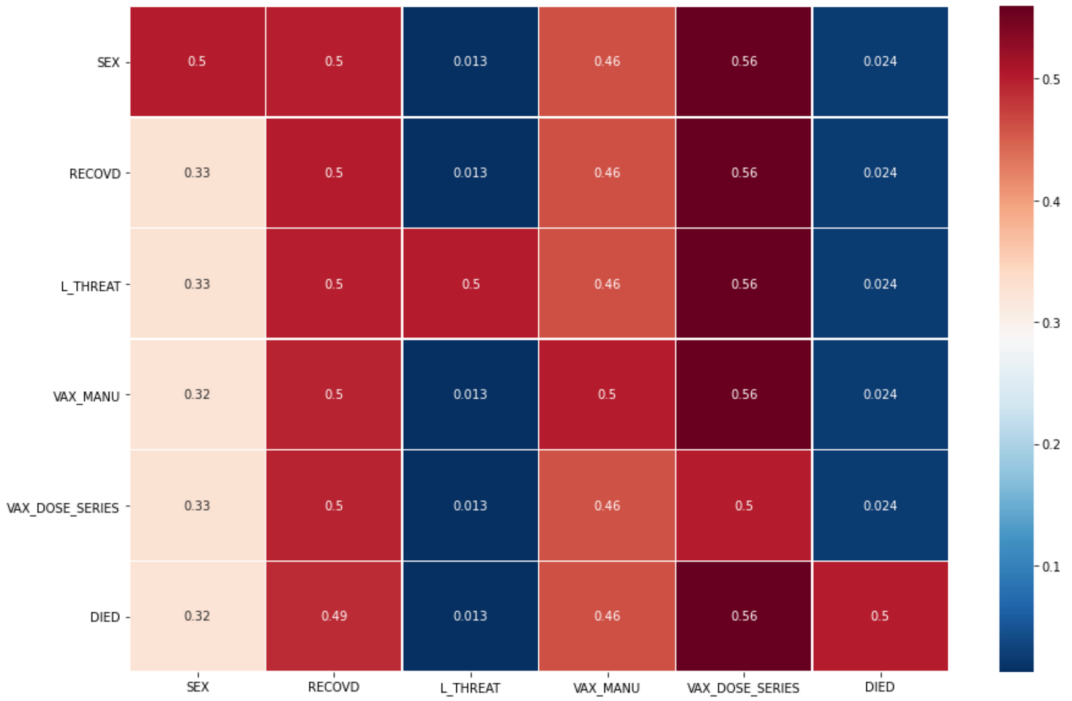
| Vaccine Symptoms | Description | | Population Features | Description | | Type of Vaccine | Description | |
|---|---|---|---|---|---|---|---|---|
| **RECOVD** | NO | 167800 | **AGE_YRS** | 0-30 | 107937 | **VAX_MANU** | MODERNA | 159109 |
| | YES | 167637 | | 30-60 | 139483 | | PFIZER | 134505 |
| **L_THREAT** | NO | 328931 | **SEX** | >60 | 88017 | **VAX_DOSE_SERIES** | 1 | 182188 |
| | YES | 6506 | | FEMALE | 238317 | | 2 | 112838 |
| | | | | MALE | 97120 | | >3 | 19813 |

## Output Features

| | | |
|---|---|---|
| **DIED*** | NO | 330229 |
| | YES | 5208 |

# Correlation Analysis

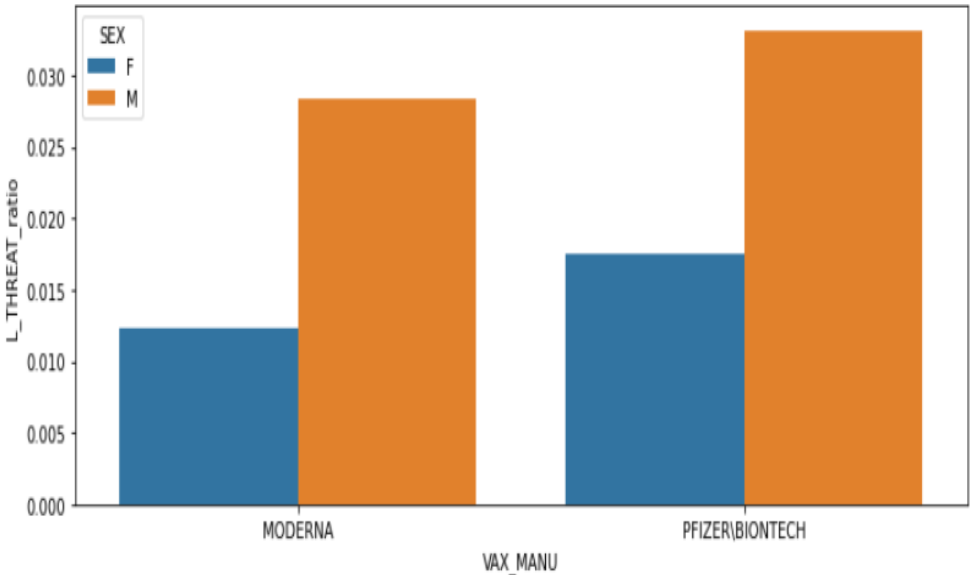
## Correlation Heat Map




- **Category variables: The Gini coefficient was used to calculate variable correlations ( row variable is the introduced condition variable )**
- **The dependent variable Gini coefficient is significantly reduced, indicating that there is a strong correlation between them.**

# Interesting Findings

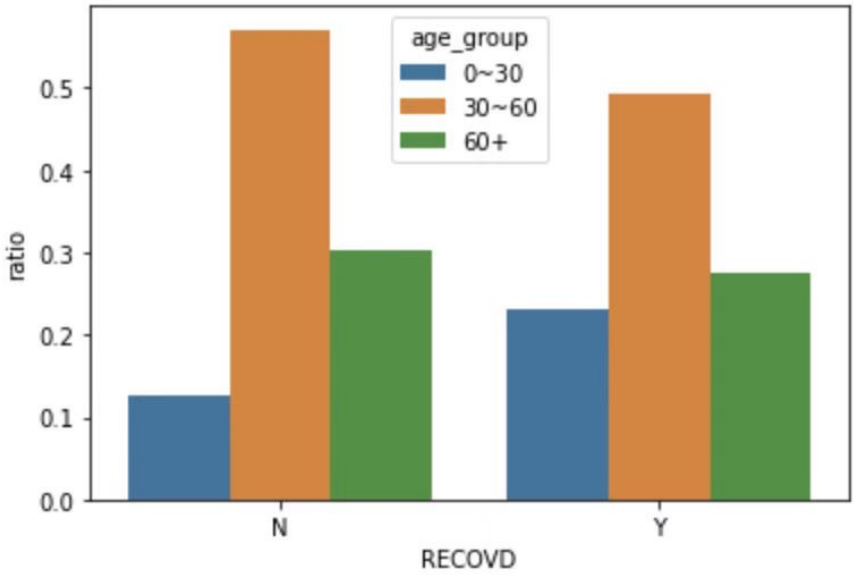
## Ratio of Having Life Threat Among Different Vaccine Manufacture




**In comparison to PFIZER, people getting vaccine with Moderna has less possibility of having life threating based on vaccination side effect.**

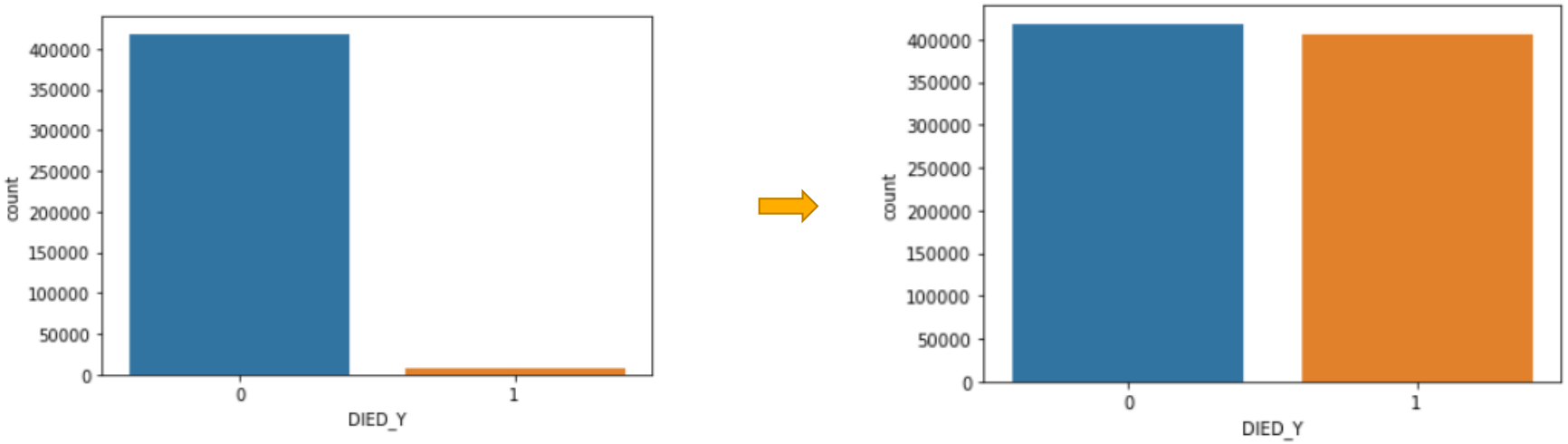
# Interesting Findings

## Ratio of Recovery from Vaccination Side Effect Among Different Age



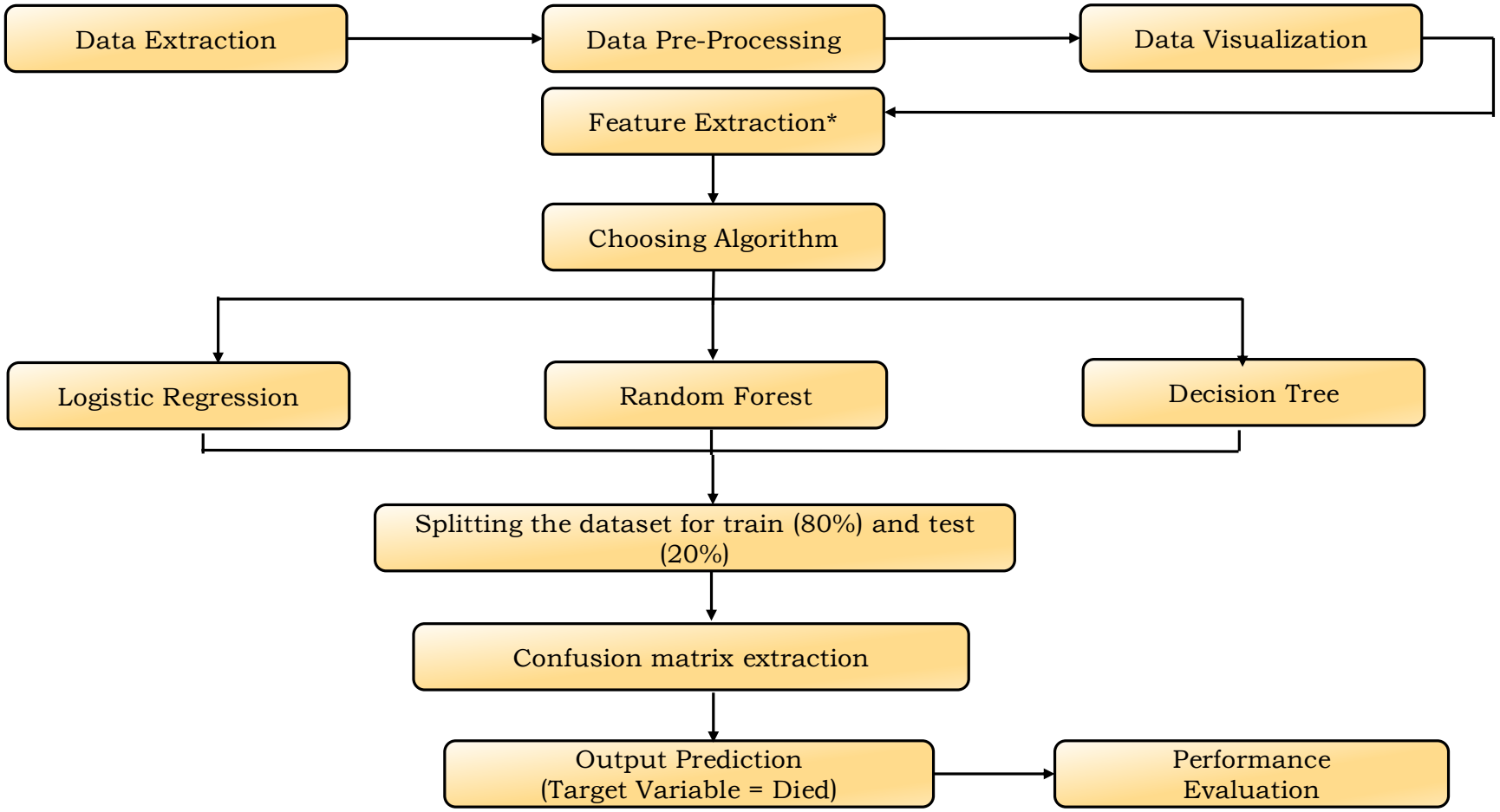

**Younger age under 30-year-old has more possibility getting recover from side effect after vaccination than other elder groups.**

# Imbalanced Data




- ✓ This problem of **imbalanced dataset** can lead to inaccurate results even when brilliant models are used to process that data.
- ✓ Use **SMOTE methods** to increase the number of cases in our dataset in a balanced way.

# Working Process



/03

# Model & Results

# Logistic Regression

| Score-Evaluating Prediction Results | | | | |
|---|---|---|---|---|
| | Precision | Recall | Accuracy | F1-Score |
| Score Result | 0.861 | 0.906 | 0.882 | 0.883 |

| Confusion Matrix | | | |
|---|---|---|---|
| | | Predict | |
| | | No Die | Die |
| Actual | No Die | 71596 | 11869 |
| | Die | 7608 | 73596 |

Our baseline model

# Decision Tree

| Grid Search | | | |
|---|---|---|---|
| | Criterion | Max_Depth | Max Samples Leaf |
| Range | 'gini' 'entropy' | 5, 10, 15, 20, 25 | 5, 10, 15 |
| Best | entropy | 25 | 5 |

| Confusion Matrix | | | |
|---|---|---|---|
| | | Predict | |
| | | Not Die | Die |
| Actual | Not Die | 72586 | 10879 |
| | Die | 11788 | 69416 |

| Score-Evaluating Prediction Results | | | | |
|---|---|---|---|---|
| | Precision | Recall | Accuracy | F1-Score |
| Score Result | 0.866 | 0.855 | 0.862 | 0.859 |

**No better than our baseline model**

# Random Forest

| Grid Search | | | | |
|---|---|---|---|---|
| | n_estimators | Max_Depth | Max_Features | Max_Samples_Split |
| Range | start =200, stop=1000,num=10 | start =10, stop=100,num=10 | 'auto','sqrt | 2,5,10 |
| Best | 911 | 80 | auto | 10 |

| Confusion Matrix | | | |
|---|---|---|---|
| | | Predict | |
| | | Not Die | Die |
| Actual | Not Die | 82230 | 13 |
| | Die | 1235 | 81191 |

| Score-Evaluating Prediction Results | | | | |
|---|---|---|---|---|
| | Precision | Recall | Accuracy | F1-Score |
| Score Result | 0.985 | 0.998 | 0.992 | 0.991 |

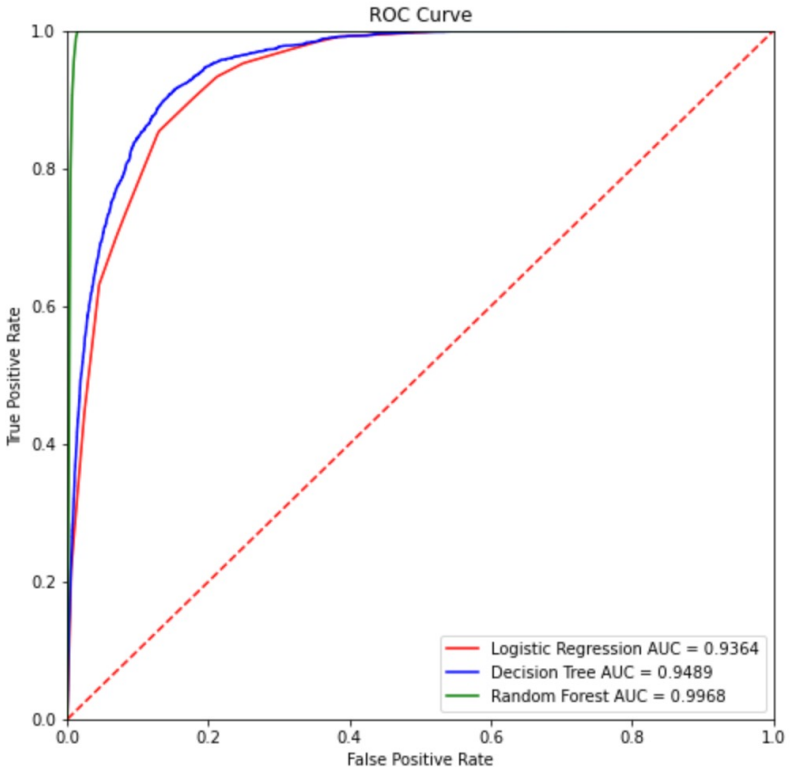
| Cross Validation (CV=5) | | | |
|---|---|---|---|
| | Mean | std | Quantile (0.0025) |
| Result | 0.992 | 0.002 | [0.9915,0.9917] |

**High CV scores and low standard means the model fits well**

/04

# Conclusion & Challenge

# Performance Evaluation

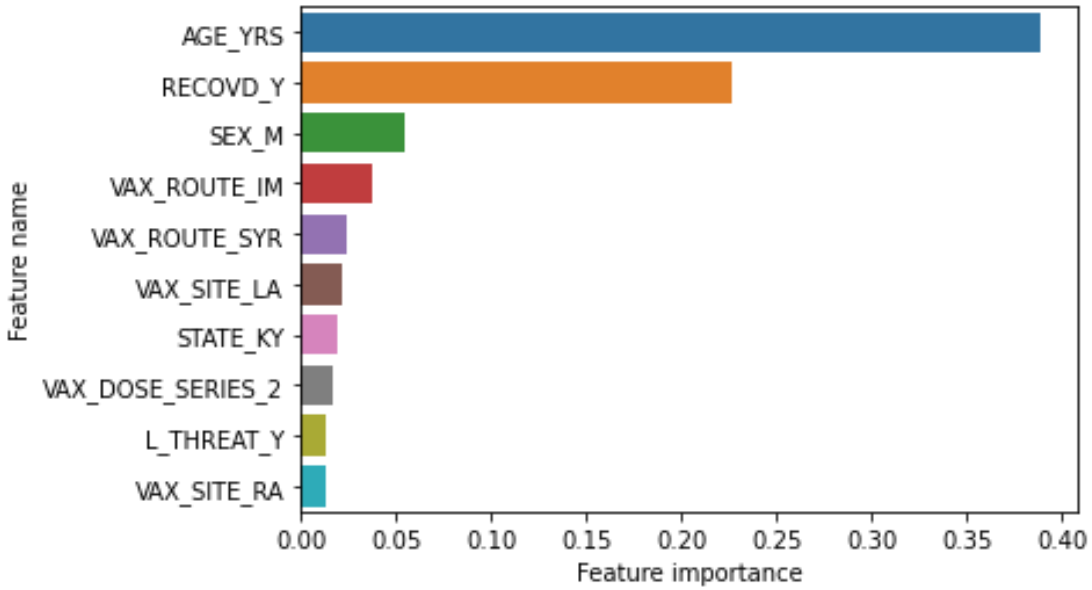



**Scores-Evaluating Prediction Results**

| No | Approach | Precision | Recall | Accuracy | F1-Score |
|---|---|---|---|---|---|
| 1 | LR | 0.861 | 0.906 | 0.883 | 0.882 |
| 2 | DT | 0.866 | 0.855 | 0.859 | 0.862 |
| 3 | RF | 0.985 | 0.998 | 0.991 | 0.992 |

- **Logistic regression is our baseline model.**
- **Random forest perform best.**

# Important Features




- **The 10 most important variables in the random forest model.**
- **AGE_YRS and ROVODE_Y are far more important than the other variables**

# Conclusion &Challenge

## Conclusion

- Low mortality rate.
- Correlation with Age.
- The Random Forest model achieves relatively better results and can be used to predict the probability of death.
- The most important variable in the Random Forest model is age.

## Challenge

- Model:
  More models should be tried, such as SVM.

  The parameters of the existing model should be adjusted in detail to avoid overfitting.
- Data:
  A lot of information & Too many missing values need to delete.

  More methods should be used to refine the dataset to obtain more features.

# Thank You