

BDA 2020/21

Datasets for projects

more details at this link:

http://bit.ly/bda2021_datasets

Heart Failure Prediction

12 variables

(age, anaemia, creatinine, diabetes, sex, etc.)

1 target

(1 patient has died, 0 has not)

- **unbalanced data set**
- **small data set**
(300 instances)

Cardiovascular diseases (CVDs) are the number 1 cause of death globally (31% of all deaths worldwide).

kaggle

<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

Google Play Store Apps

14 variables

(name, category, user rating, reviews, dimension of the app, downloads, price, age group, genre, date, version, ...)

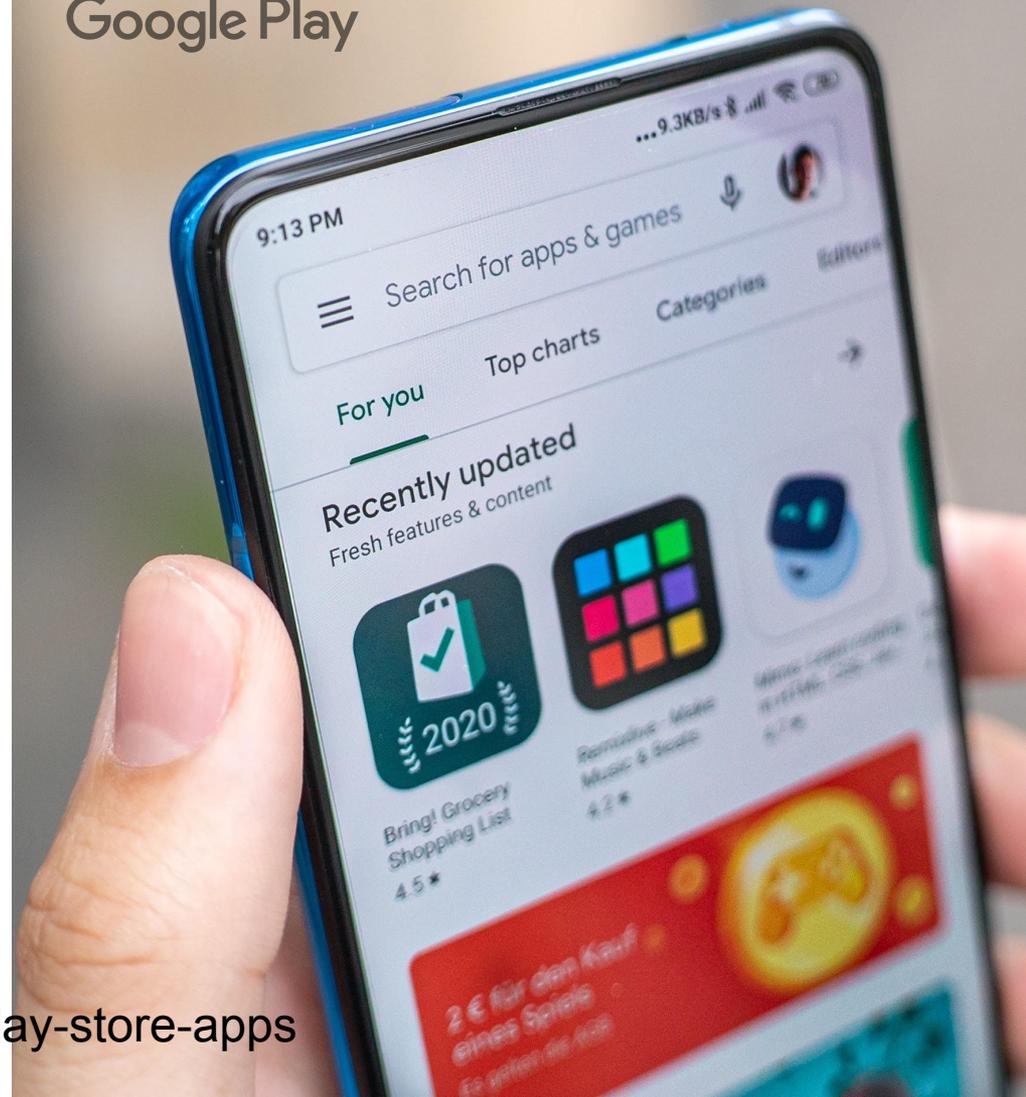
target: success of the app (e.g., the rating)

<https://www.kaggle.com/lava18/google-play-store-apps>



Google Play

kaggle



Red Wine Quality

11 variables based on physicochemical tests

fixed acidity, acidity, citric acid, sugar, chlorides, free and total sulfur dioxide, density, pH, sulphates, alcohol

1 target: *quality* score in $[0, 10]$

ordered classes
unbalanced dataset



<https://archive.ics.uci.edu/ml/datasets/wine+quality>

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

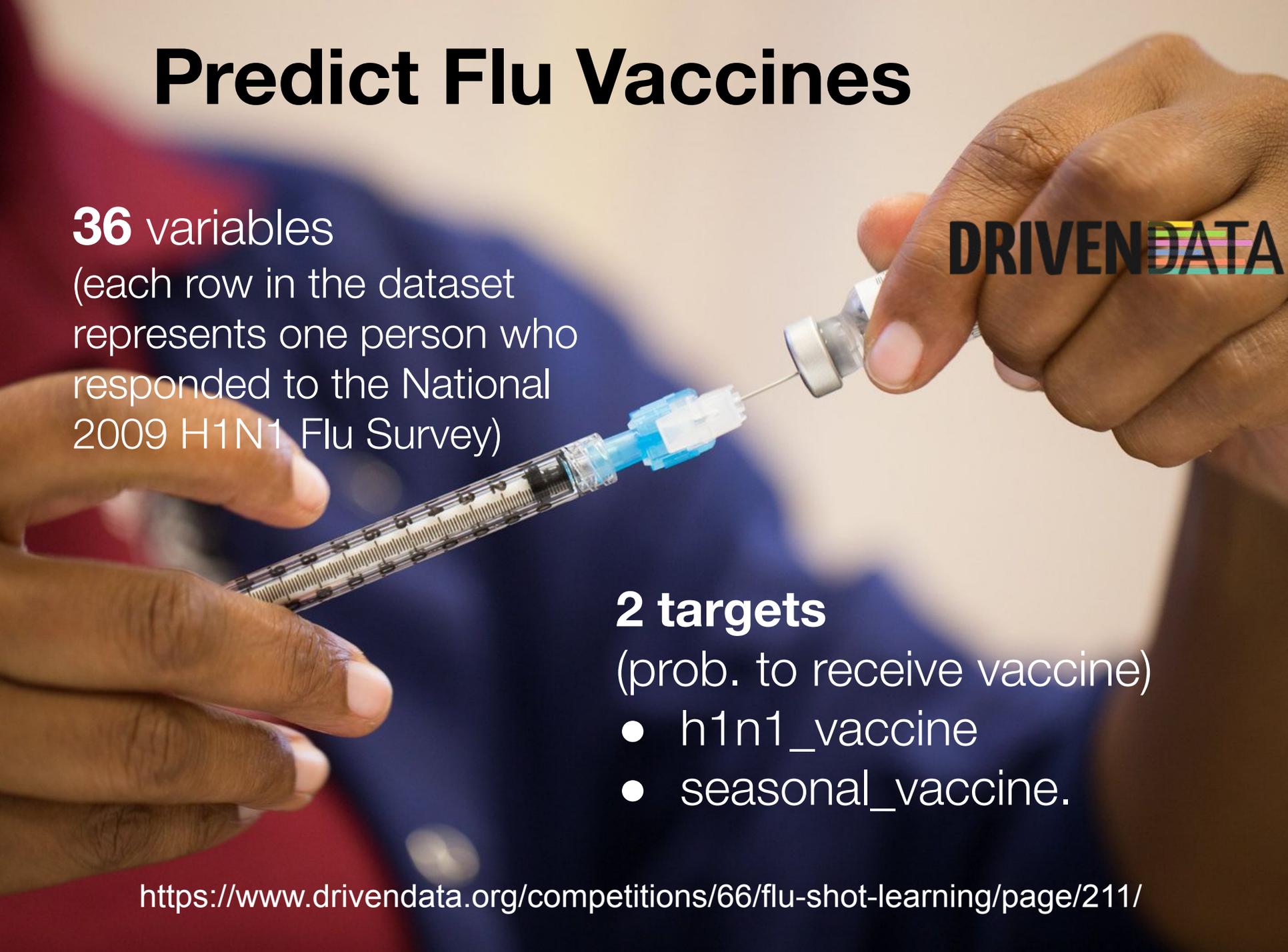
Modeling Earthquake Damage

39 variables
(information on the
buildings' structure)

DRIVENDATA

1 target: severity of
damage (low, medium,
high)

Predict Flu Vaccines



36 variables

(each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey)

DRIVEN DATA

2 targets

(prob. to receive vaccine)

- h1n1_vaccine
- seasonal_vaccine.

Pump it Up: Data Mining the Water Table

40 variables

(amount of water, funder, altitude, position, year, kind of extraction, management, costs, water quality)

1 target: functional state
(functional, needs repair, non functional)

Predicting Disease Spread

DRIVENDATA

20+ variables
(temperature, precipitation,
humidity, vegetation, and
more.)

1 target: total
number of cases
for each (city, year,
weekofyear).