Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

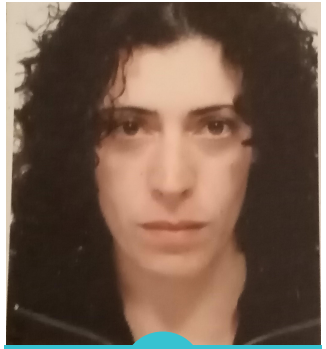
Big Data Analytics
A.A. 2020/21

MaLuCS
Luca Corbucci
Cinzia Lestini
Marco Giuseppe Marino
Simone Rossi

MaLuCS TEAM

LUCA CORBUCCI

Master Degree in Computer Science

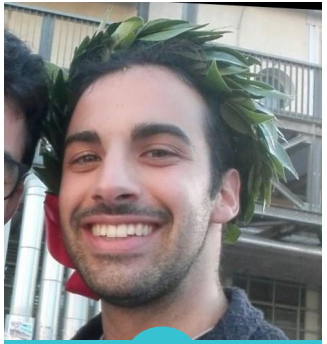


CINZIA LESTINI

Master Degree Data Science and Business Informatics

MARCO GIUSEPPE MARINO

Master Degree in Computer Science



SIMONE ROSSI

Master Degree in Computer Science

INTRODUCTION

DATASET DESCRIPTION – KEY NUMBERS

26,707

Rows in training dataset

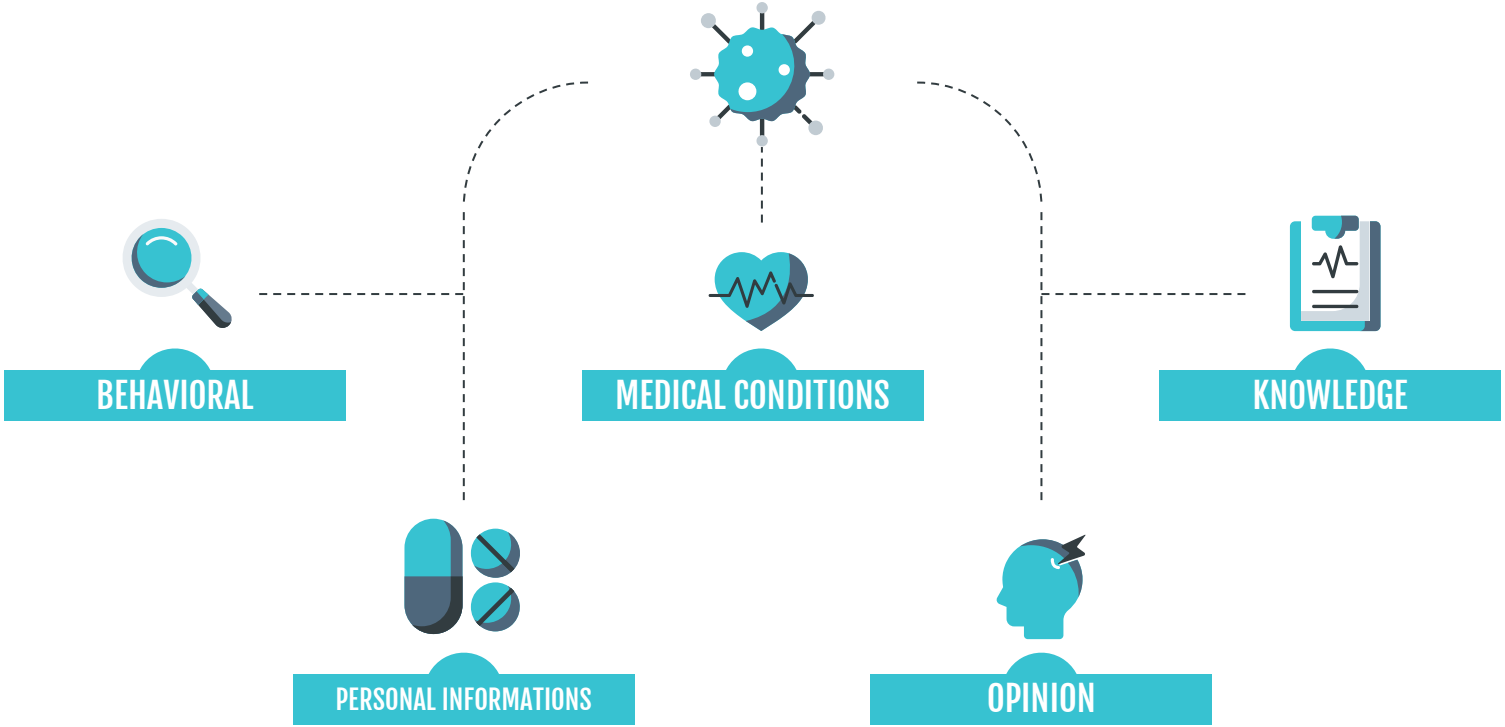
2

Target variables

36

Features

DATASET DESCRIPTION – FEATURES



DATASET DESCRIPTION – FEATURES

Ordinal features

(opinion_h1n1_vacc_effective,
opinion_h1n1_risk,
opinion_h1n1_sick_from_vacc,
opinion_seas_vacc_effective,
.....)

Binary Features

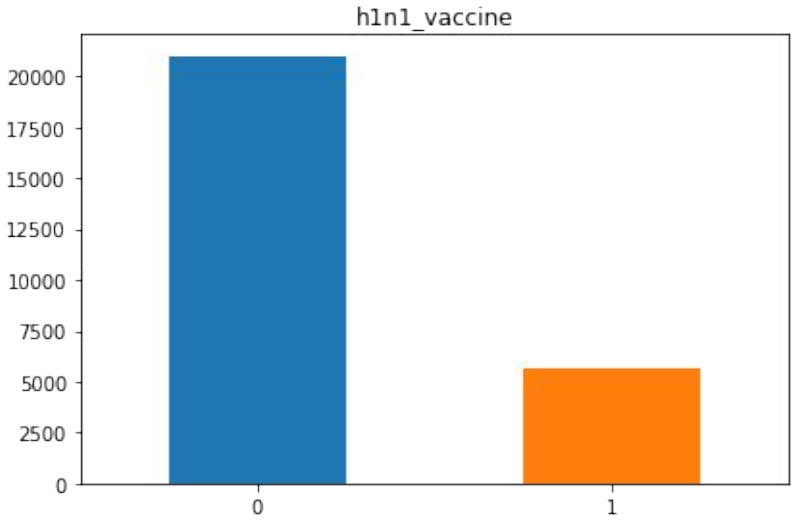
(h1n1_knowledge....
behavioral_face_mask..
doctor_recc_H1N1.....
chronic_med_condition..)

Nominal features

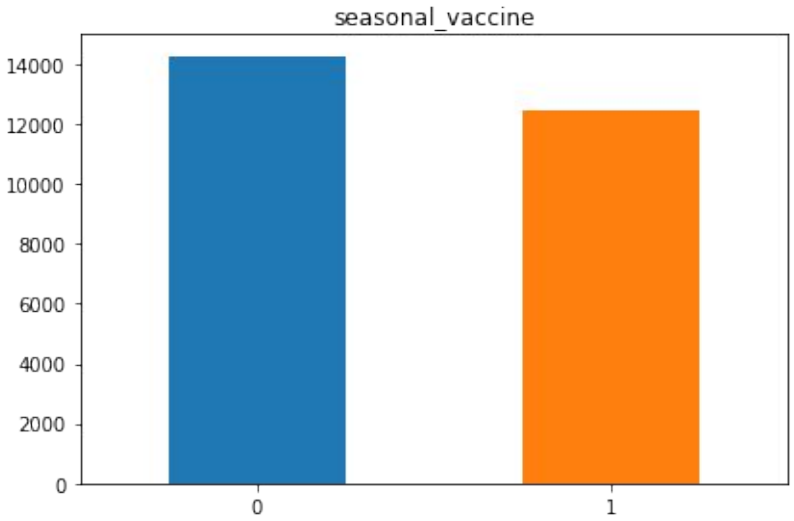
(age_group, education, race, sex,
income_poverty, marital_status,
employment_status .....)

TARGET VARIABLES

H1N1 VACCINE



SEASONAL VACCINE



DATA CLEANING

DATA CLEANING

We analysed all the columns to find and remove the missing values

MISSING VALUES

We checked the dataset to be sure that we do not have outliers and repeated rows

DATA CONSISTENCY

We integer encoded the categorical variables in our dataset

VARIABLES ENCODING

MISSING VALUES

How many Missing Values?

In our dataset we found 30 features with missing values

Mode

We filled some missing values with the mode of the column

Drop Features

We dropped 3 columns where the missing values were half of the column

Grouped Mode

We filled some missing values grouping on a correlated features and then we used the mode

DATA CONSISTENCY AND VARIABLES ENCODING

Data Consistency

- Outliers
- “Strange” Values
- Duplicated Rows

Variables Encoding

- We encoded all the categorical features to numbers.
- This will be useful for the computation of the correlation.

DATA EXPLORATION

DATA EXPLORATION

We show all the possible values of the data with a bar plot for each column.

DISTRIBUTION

The features conditional distribution with respect to the target variables.

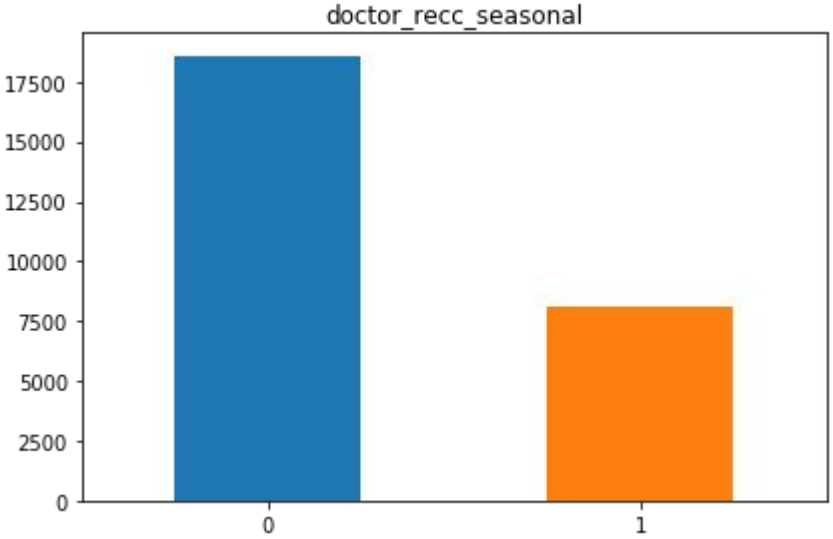
CONDITIONAL HISTOGRAMS

The Pearson's linear correlation between variables.

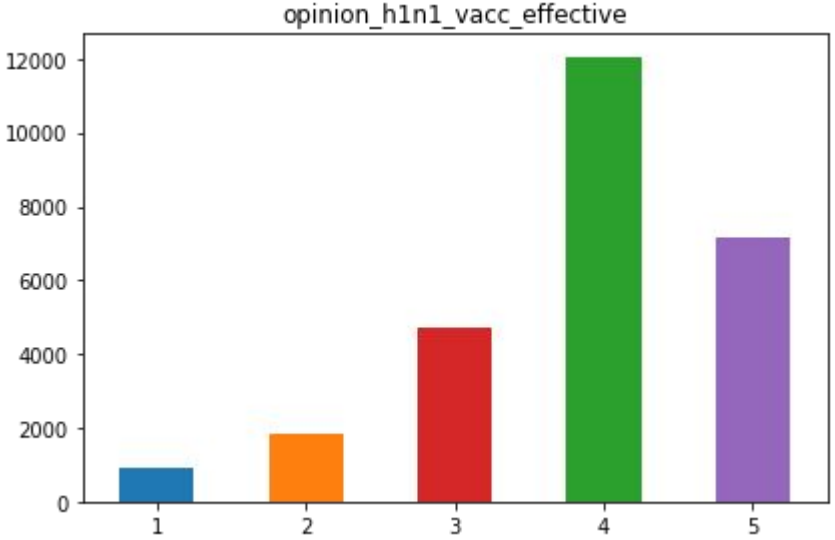
CORRELATIONS

DATA DISTRIBUTION

BINARY

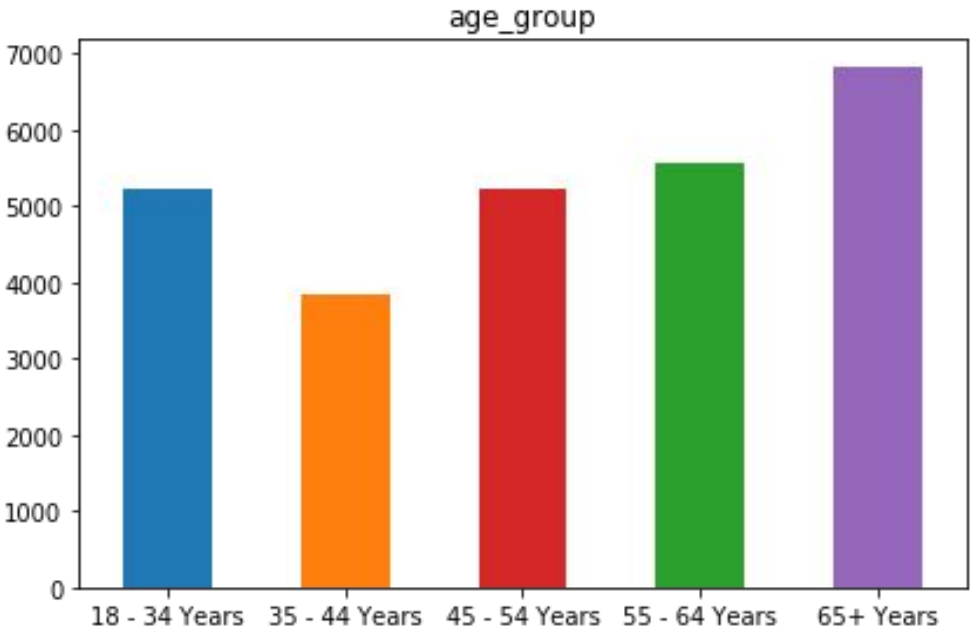


ORDINAL

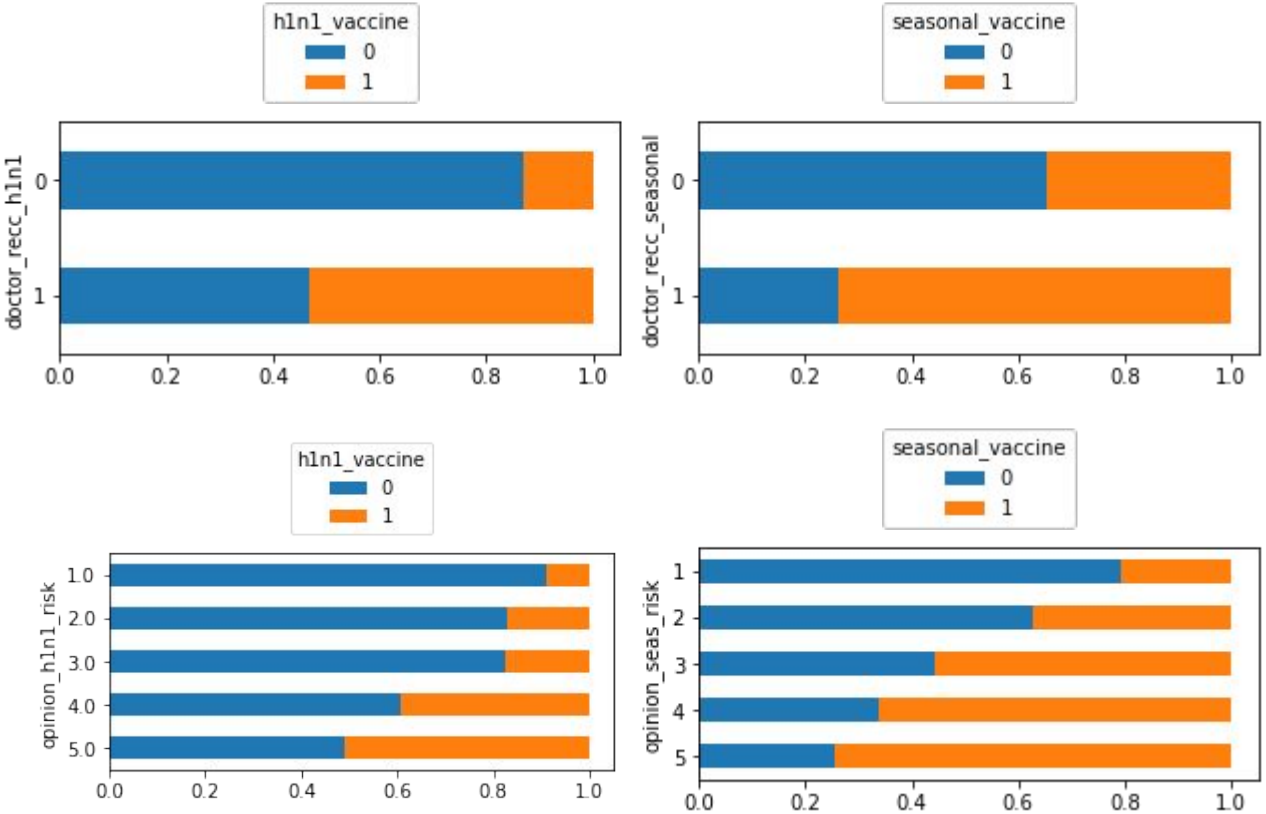


DATA DISTRIBUTION

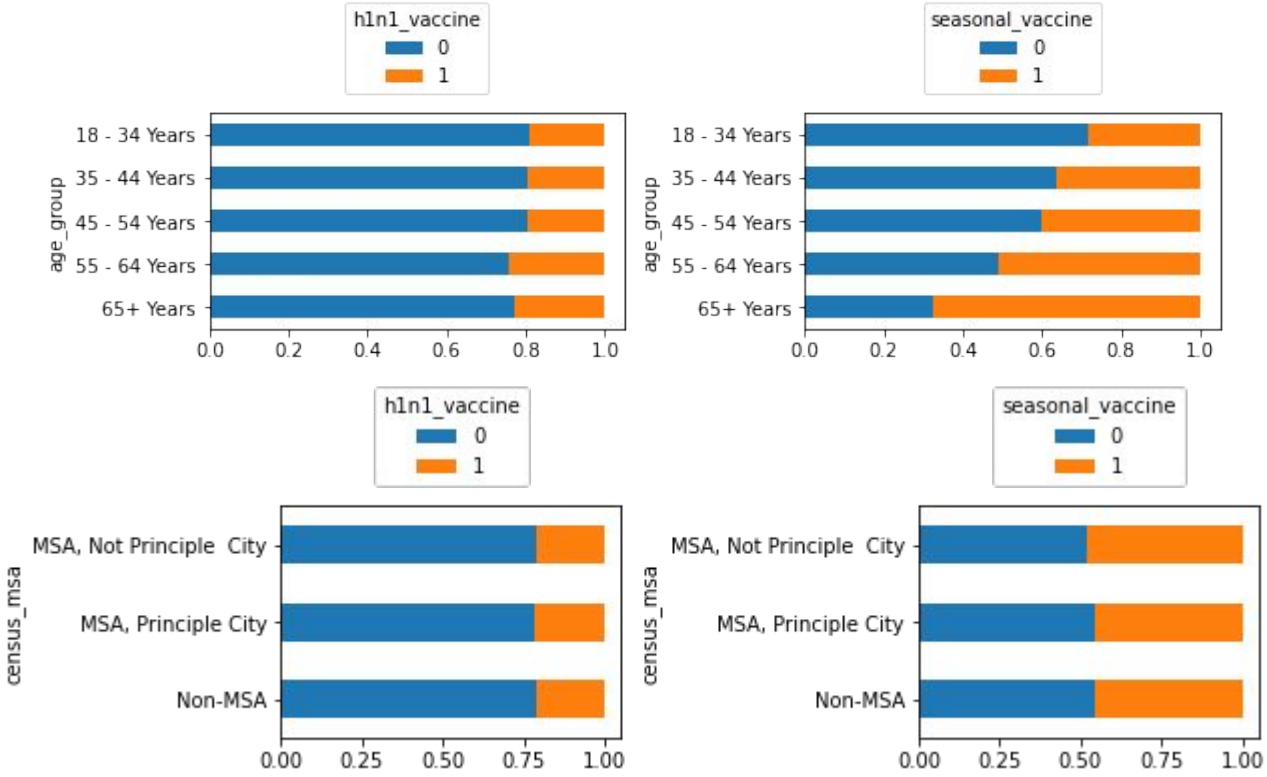
NOMINAL



CONDITIONAL HISTOGRAMS



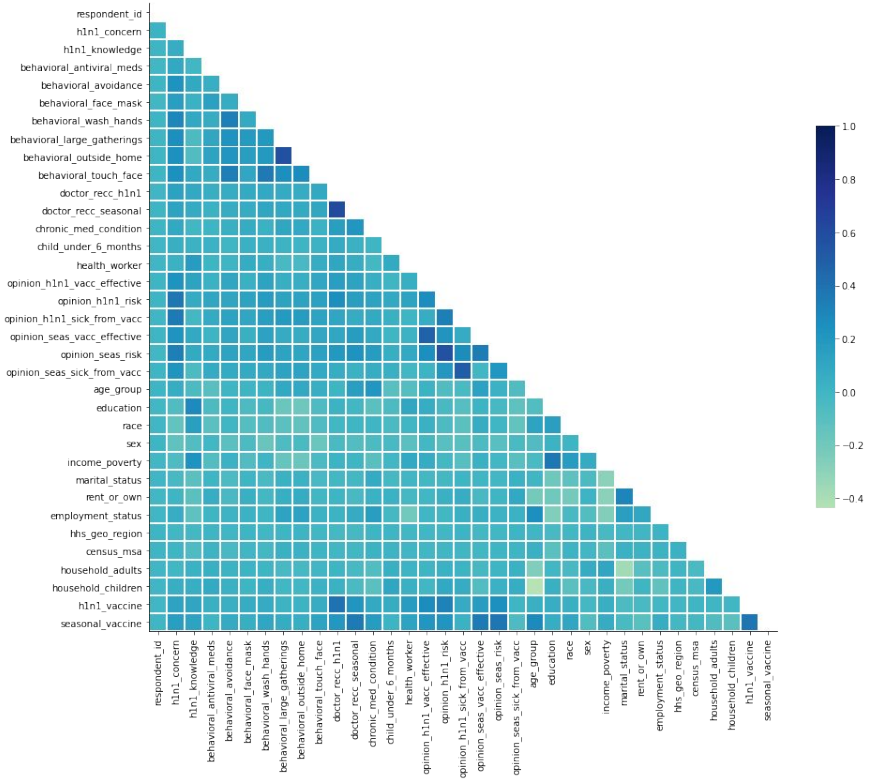
CONDITIONAL HISTOGRAMS



CORRELATIONS

CORRELATION ANALYSIS

- weak correlation: c ≥ 20
- moderate correlation: c > 30
- strong correlation: c > 40
- very strong correlation: c > 70



CORRELATION BETWEEN FEATURES

h1n1_concern

Weakly correlated to behavior and opinion variables.

Behavior

General moderate/weak correlation, strong between behavioral_outside_home and behavioral_large_gatherings

Opinion

General moderate/weak correlation, but effectivity and getting sick from the vaccine are not related.

Social and economic

Positive and negative correlation.

CORRELATION OF TARGET

H1N1

1. doctor_recc_h1n1 : 0.39
2. opinion_h1n1_risk: 0.32
3. opinion_h1n1_vacc_effective: 0.26

doctor_recc_h1n1	correlation likelihood
0	0.13
1	0.53

Seasonal flu

1. opinion_seas_risk : 0.38
2. doctor_recc_seasonal: 0.36
3. opinion_seas_vacc_effective: 0.35

opinion_seas_risk	correlation likelihood
1	0.20
2	0.37
3	0.55
4	0.66
5	0.74

DATA TRANSFORMATION

DATA TRANSFORMATION

Behaviour

We created a single feature summing all the behavioural features.

Family Size

We created a new feature for the Family Size of the respondent

Variable Elimination

We deleted some features that are correlated with others in the dataset.

Useless Features

We think that other features will be useless during the classification but we did not removed them because are not correlated with others.

CONCLUSIONS

This first Milestone led use to the know our dataset and to propose some possible changes that we will exploit during the Classification Phase.

THANKS!

Do you have any questions?

REFERENCES

- U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.
- Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines