Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

Caroline Gunawan, Clarisa Angeline, Yohan Chandrasukmana

5/23/2022

Contents

1 Data Loading **2**

2 Data Cleaning **7**

2.1 Imputasi Variabel `health_insurance` 8

2.2 Imputasi Variabel Dependen Lainnya 9

2.3 Type-casting Variabel Independen 9

3 Analisa Variabel Independen **10**

4 Train-Test Split pada Train Set **14**

5 Model **14**

5.1 Logistic Regression 15

5.2 Naive Bayes 23

5.3 GBM 33

6 Hasil Akhir **43**

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tictoc)
```

1 Data Loading

```
train_labels <- read.csv("training_set_labels.csv")
train_features <- read.csv("training_set_features.csv")
test_features <- read.csv("test_set_features.csv")
```

```
head(train_labels)
```

```
##   respondent_id h1n1_vaccine seasonal_vaccine
## 1             0            0               0
## 2             1            0               1
## 3             2            0               0
## 4             3            0               1
## 5             4            0               0
## 6             5            0               0
```

```
head(train_features)
```

```
##   respondent_id h1n1_concern h1n1_knowledge behavioral_antiviral_meds
## 1             0            1              0                         0
## 2             1            3              2                         0
## 3             2            1              1                         0
## 4             3            1              1                         0
## 5             4            2              1                         0
## 6             5            3              1                         0
##   behavioral_avoidance behavioral_face_mask behavioral_wash_hands
## 1                    0                    0                     0
## 2                    1                    0                     1
## 3                    1                    0                     0
## 4                    1                    0                     1
## 5                    1                    0                     1
## 6                    1                    0                     1
##   behavioral_large_gatherings behavioral_outside_home behavioral_touch_face
## 1                           0                       1                     1
## 2                           0                       1                     1
## 3                           0                       0                     0
## 4                           1                       0                     0
## 5                           1                       0                     1
## 6                           0                       0                     1
##   doctor_recc_h1n1 doctor_recc_seasonal chronic_med_condition
## 1                0                    0                     0
## 2                0                    0                     0
## 3               NA                   NA                     1
## 4                0                    1                     1
## 5                0                    0                     0
## 6                0                    1                     0
##   child_under_6_months health_worker health_insurance
```

```
## 1                    0            0               1
## 2                    0            0               1
## 3                    0            0              NA
## 4                    0            0              NA
## 5                    0            0              NA
## 6                    0            0              NA
##   opinion_h1n1_vacc_effective opinion_h1n1_risk opinion_h1n1_sick_from_vacc
## 1                           3                1                           2
## 2                           5                4                           4
## 3                           3                1                           1
## 4                           3                3                           5
## 5                           3                3                           2
## 6                           5                2                           1
##   opinion_seas_vacc_effective opinion_seas_risk opinion_seas_sick_from_vacc
## 1                           2                1                           2
## 2                           4                2                           4
## 3                           4                1                           2
## 4                           5                4                           1
## 5                           3                1                           4
## 6                           5                4                           4
##       age_group        education  race    sex            income_poverty
## 1 55 - 64 Years       < 12 Years White Female             Below Poverty
## 2 35 - 44 Years         12 Years White   Male             Below Poverty
## 3 18 - 34 Years College Graduate White   Male <= $75,000, Above Poverty
## 4     65+ Years         12 Years White Female             Below Poverty
## 5 45 - 54 Years     Some College White Female <= $75,000, Above Poverty
## 6     65+ Years         12 Years White   Male <= $75,000, Above Poverty
##   marital_status rent_or_own  employment_status hhs_geo_region
## 1    Not Married         Own Not in Labor Force       oxchjgsf
## 2    Not Married        Rent           Employed       bhuqouqj
## 3    Not Married         Own           Employed       qufhixun
## 4    Not Married        Rent Not in Labor Force       lrircsnp
## 5        Married         Own           Employed       qufhixun
## 6        Married         Own           Employed       atmpeygn
##                 census_msa household_adults household_children
## 1                   Non-MSA               0                  0
## 2 MSA, Not Principle  City               0                  0
## 3 MSA, Not Principle  City               2                  0
## 4      MSA, Principle City               0                  0
## 5 MSA, Not Principle  City               1                  0
## 6      MSA, Principle City               2                  3
##   employment_industry employment_occupation
## 1
## 2            pxcmvdjn              xgwztkwe
## 3            rucpziij              xtkaffoo
## 4
## 5            wxleyezf              emcorrxb
## 6            saaquncn              vlluhbov
```

```
head(test_features)
```

```
##   respondent_id h1n1_concern h1n1_knowledge behavioral_antiviral_meds
## 1         26707            2              2                         0
## 2         26708            1              1                         0
```

```
## 3         26709           2             2                       0
## 4         26710           1             1                       0
## 5         26711           3             1                       1
## 6         26712           2             2                       0
##   behavioral_avoidance behavioral_face_mask behavioral_wash_hands
## 1                    1                    0                     1
## 2                    0                    0                     0
## 3                    0                    1                     1
## 4                    0                    0                     0
## 5                    1                    0                     1
## 6                    1                    0                     1
##   behavioral_large_gatherings behavioral_outside_home behavioral_touch_face
## 1                           1                       0                     1
## 2                           0                       0                     0
## 3                           1                       1                     1
## 4                           0                       0                     0
## 5                           1                       1                     1
## 6                           1                       0                     1
##   doctor_recc_h1n1 doctor_recc_seasonal chronic_med_condition
## 1                0                    0                    0
## 2                0                    0                    0
## 3                0                    0                    0
## 4                1                    1                    1
## 5                0                    0                    0
## 6                0                    0                    0
##   child_under_6_months health_worker health_insurance
## 1                    0             0                1
## 2                    0             0                0
## 3                    0             0               NA
## 4                    0             0                1
## 5                    0             1                1
## 6                    0             1                1
##   opinion_h1n1_vacc_effective opinion_h1n1_risk opinion_h1n1_sick_from_vacc
## 1                           5                 1                           1
## 2                           4                 1                           1
## 3                           5                 4                           2
## 4                           4                 2                           2
## 5                           5                 2                           4
## 6                           4                 4                           1
##   opinion_seas_vacc_effective opinion_seas_risk opinion_seas_sick_from_vacc
## 1                           5                 1                           1
## 2                           4                 1                           1
## 3                           5                 4                           4
## 4                           4                 4                           2
## 5                           4                 4                           2
## 6                           5                 5                           1
##       age_group        education     race    sex            income_poverty
## 1 35 - 44 Years College Graduate Hispanic Female                > $75,000
## 2 18 - 34 Years         12 Years    White   Male              Below Poverty
## 3 55 - 64 Years College Graduate    White   Male                > $75,000
## 4     65+ Years         12 Years    White Female <= $75,000, Above Poverty
## 5 35 - 44 Years         12 Years    Black Female <= $75,000, Above Poverty
## 6 45 - 54 Years College Graduate    White Female                > $75,000
##   marital_status rent_or_own  employment_status hhs_geo_region
```

```
## 1    Not Married        Rent           Employed       mlyzmhmf
## 2    Not Married        Rent           Employed       bhuqouqj
## 3        Married         Own           Employed       lrircsnp
## 4        Married         Own Not in Labor Force       lrircsnp
## 5    Not Married         Own           Employed       lzgpxyit
## 6    Not Married         Own           Employed       mlyzmhmf
##                census_msa household_adults household_children
## 1 MSA, Not Principle  City               1                  0
## 2                 Non-MSA               3                  0
## 3                 Non-MSA               1                  0
## 4 MSA, Not Principle  City               1                  0
## 5                 Non-MSA               0                  1
## 6      MSA, Principle City               0                  2
##   employment_industry employment_occupation
## 1            atmlpfrs              hfxkjkmi
## 2            atmlpfrs              xqwwgdyp
## 3            nduyfdeo              pvmttkik
## 4
## 5            fcxhlnwr              mxkfnird
## 6            fcxhlnwr              cmhcxjea
```

```
# Menggabungkan Training Labels dan Features
train <- merge(train_labels, train_features, by="respondent_id")
head(train)
```

```
##   respondent_id h1n1_vaccine seasonal_vaccine h1n1_concern h1n1_knowledge
## 1             0            0                0           1              0
## 2             1            0                1           3              2
## 3             2            0                0           1              1
## 4             3            0                1           1              1
## 5             4            0                0           2              1
## 6             5            0                0           3              1
##   behavioral_antiviral_meds behavioral_avoidance behavioral_face_mask
## 1                         0                    0                   0
## 2                         0                    1                   0
## 3                         0                    1                   0
## 4                         0                    1                   0
## 5                         0                    1                   0
## 6                         0                    1                   0
##   behavioral_wash_hands behavioral_large_gatherings behavioral_outside_home
## 1                     0                           0                       1
## 2                     1                           0                       1
## 3                     0                           0                       0
## 4                     1                           1                       0
## 5                     1                           1                       0
## 6                     1                           0                       0
##   behavioral_touch_face doctor_recc_h1n1 doctor_recc_seasonal
## 1                     1               0                   0
## 2                     1               0                   0
## 3                     0              NA                  NA
## 4                     0               0                   1
## 5                     1               0                   0
## 6                     1               0                   1
##   chronic_med_condition child_under_6_months health_worker health_insurance
```

```
## 1                     0                   0            0                1
## 2                     0                   0            0                1
## 3                     1                   0            0               NA
## 4                     1                   0            0               NA
## 5                     0                   0            0               NA
## 6                     0                   0            0               NA
##   opinion_h1n1_vacc_effective opinion_h1n1_risk opinion_h1n1_sick_from_vacc
## 1                           3                1                           2
## 2                           5                4                           4
## 3                           3                1                           1
## 4                           3                3                           5
## 5                           3                3                           2
## 6                           5                2                           1
##   opinion_seas_vacc_effective opinion_seas_risk opinion_seas_sick_from_vacc
## 1                           2                1                           2
## 2                           4                2                           4
## 3                           4                1                           2
## 4                           5                4                           1
## 5                           3                1                           4
## 6                           5                4                           4
##       age_group        education  race    sex            income_poverty
## 1 55 - 64 Years       < 12 Years White Female             Below Poverty
## 2 35 - 44 Years         12 Years White   Male             Below Poverty
## 3 18 - 34 Years College Graduate White   Male <= $75,000, Above Poverty
## 4     65+ Years         12 Years White Female             Below Poverty
## 5 45 - 54 Years     Some College White Female <= $75,000, Above Poverty
## 6     65+ Years         12 Years White   Male <= $75,000, Above Poverty
##   marital_status rent_or_own  employment_status hhs_geo_region
## 1    Not Married         Own Not in Labor Force       oxchjgsf
## 2    Not Married        Rent           Employed       bhuqouqj
## 3    Not Married         Own           Employed       qufhixun
## 4    Not Married        Rent Not in Labor Force       lrircsnp
## 5        Married         Own           Employed       qufhixun
## 6        Married         Own           Employed       atmpeygn
##                census_msa household_adults household_children
## 1                  Non-MSA                0                  0
## 2 MSA, Not Principle  City                0                  0
## 3 MSA, Not Principle  City                2                  0
## 4      MSA, Principle City                0                  0
## 5 MSA, Not Principle  City                1                  0
## 6      MSA, Principle City                2                  3
##   employment_industry employment_occupation
## 1
## 2            pxcmvdjn              xgwztkwe
## 3            rucpziij              xtkaffoo
## 4
## 5            wxleyezf              emcorrxb
## 6            saaquncn              vlluhbov
```

```r
# Mengubah respondent_id menjadi index setiap entry
rownames(train) <- train$respondent_id
rownames(test_features) <- test_features$respondent_id

# Menghapus respondent_id
```

```
train = train[,-1]
test = test_features[,-1]
```

2 Data Cleaning

```
# Mengecek proporsi missing values setiap kolom.
dfmissing_train = data.frame("NA"=sapply(train,
                                         function(x) sum(is.na(x))/dim(train)[1]))
arrange(dfmissing_train, desc(NA.))
```

```
##                                    NA.
## health_insurance            0.459579885
## doctor_recc_h1n1            0.080877673
## doctor_recc_seasonal        0.080877673
## chronic_med_condition       0.036357509
## child_under_6_months        0.030703561
## health_worker               0.030104467
## opinion_seas_sick_from_vacc 0.020107088
## opinion_seas_risk           0.019245891
## opinion_seas_vacc_effective 0.017298836
## opinion_h1n1_sick_from_vacc 0.014790130
## opinion_h1n1_vacc_effective 0.014640356
## opinion_h1n1_risk           0.014528026
## household_adults            0.009323398
## household_children          0.009323398
## behavioral_avoidance        0.007788220
## behavioral_touch_face       0.004792751
## h1n1_knowledge              0.004343431
## h1n1_concern                0.003444790
## behavioral_large_gatherings 0.003257573
## behavioral_outside_home     0.003070356
## behavioral_antiviral_meds   0.002658479
## behavioral_wash_hands       0.001572621
## behavioral_face_mask        0.000711424
## h1n1_vaccine                0.000000000
## seasonal_vaccine            0.000000000
## age_group                   0.000000000
## education                   0.000000000
## race                        0.000000000
## sex                         0.000000000
## income_poverty              0.000000000
## marital_status              0.000000000
## rent_or_own                 0.000000000
## employment_status           0.000000000
## hhs_geo_region              0.000000000
## census_msa                  0.000000000
## employment_industry         0.000000000
## employment_occupation       0.000000000
```

```
dfmissing_test = data.frame("NA"=sapply(test,
                                        function(x) sum(is.na(x))/dim(test)[1]))
arrange(dfmissing_test, desc(NA.))
```

```
##                                    NA.
## health_insurance            0.4578403475
## doctor_recc_h1n1            0.0808746443
## doctor_recc_seasonal        0.0808746443
## chronic_med_condition       0.0348959113
## child_under_6_months        0.0304403175
## health_worker               0.0295417103
## opinion_seas_sick_from_vacc 0.0195072637
## opinion_seas_risk           0.0186835405
## opinion_seas_vacc_effective 0.0169237682
## opinion_h1n1_vacc_effective 0.0149019021
## opinion_h1n1_risk           0.0142279467
## opinion_h1n1_sick_from_vacc 0.0140407369
## household_adults            0.0084244421
## household_children          0.0084244421
## behavioral_avoidance        0.0079751385
## behavioral_touch_face       0.0047925715
## h1n1_knowledge              0.0045679197
## h1n1_concern                0.0031825670
## behavioral_outside_home     0.0030702411
## behavioral_antiviral_meds   0.0029579152
## behavioral_large_gatherings 0.0026958215
## behavioral_wash_hands       0.0014976786
## behavioral_face_mask        0.0007113973
## age_group                   0.0000000000
## education                   0.0000000000
## race                        0.0000000000
## sex                         0.0000000000
## income_poverty              0.0000000000
## marital_status              0.0000000000
## rent_or_own                 0.0000000000
## employment_status           0.0000000000
## hhs_geo_region              0.0000000000
## census_msa                  0.0000000000
## employment_industry         0.0000000000
## employment_occupation       0.0000000000
```

Variabel `health_insurance` memiliki proporsi missing values yang secara relatif besar di train dan test set.

2.1 Imputasi Variabel `health_insurance`

Dikarenakan variabel `health_insurance` memiliki proporsi missing values yang besar, akan diimputasikan sebuah nilai "-1" untuk missing values, yang mana `health_insurance` merupakan variabel biner.

```
train$health_insurance[is.na(train$health_insurance)] = -1
test$health_insurance[is.na(test$health_insurance)] = -1
```

```
# Pengecekan Missing Value setelah Imputasi
test$health_insurance[is.na(test$health_insurance)] = -1
c("Train Missing Value" = sum(is.na(train$health_insurance)),
  "Test Missing Value" = sum(is.na(test$health_insurance)))
```

```
## Train Missing Value  Test Missing Value
##                   0                   0
```

2.2 Imputasi Variabel Dependen Lainnya

```
missing_train = colnames(train)[colSums(is.na(train)) > 0]
missing_test = colnames(test)[colSums(is.na(test)) > 0]
ifelse(sum(missing_train==missing_test)==length(missing_train),
       print("Same Columns w/ Missing Values"),
       print("Different Columns w/ Missing Values"))
```

```
## [1] "Same Columns w/ Missing Values"
```

```
## [1] "Same Columns w/ Missing Values"
```

```
for(missing in missing_test){
  col = which(colnames(train)==missing)
  train[is.na(train[, col]), col] = median(train[, col], na.rm=T)

  col = which(colnames(test)==missing)
  test[is.na(test[, col]), col] = median(test[, col], na.rm=T)
}
```

2.3 Type-casting Variabel Independen

```
binary <- c('behavioral_antiviral_meds',
            'behavioral_avoidance',
            'behavioral_face_mask',
            'behavioral_wash_hands',
            'behavioral_large_gatherings',
            'behavioral_outside_home',
            'behavioral_touch_face',
            'doctor_recc_h1n1',
            'doctor_recc_seasonal',
            'chronic_med_condition',
            'child_under_6_months',
            'health_worker',
            'health_insurance', 'h1n1_vaccine')
```

```
# Type-casting tipe data variabel kategorikal ke dalam tipe factor
categorical <- c('age_group',
                 'education',
```

```
                  'race',
                  'sex',
                  'income_poverty',
                  'marital_status',
                  'rent_or_own',
                  'employment_status',
                  'hhs_geo_region',
                  'census_msa',
                  'employment_industry',
                  'employment_occupation')
train[, categorical] = lapply(train[, categorical], factor)
test[, categorical] = lapply(test[, categorical], factor)
```

```
numerical = names(test[, !names(test)%in%c(categorical,binary)])
```

3 Analisa Variabel Independen

```
library(ggplot2) # Plotting

# Target Variabel akan diubah sebagai kategorikal secara sementara
# untuk kepentingan visualisasi.
train$h1n1_vaccine = as.factor(train$h1n1_vaccine)
train$seasonal_vaccine = as.factor(train$seasonal_vaccine)
```

```
lapply(categorical,
       function(x) ggplot(train, aes(h1n1_vaccine, ..count..))
       + geom_bar(aes_string(fill=x), position="dodge"))
```

```
## [[1]]

##
## [[2]]

##
## [[3]]

##
## [[4]]

##
## [[5]]

##
## [[6]]

##
## [[7]]
```

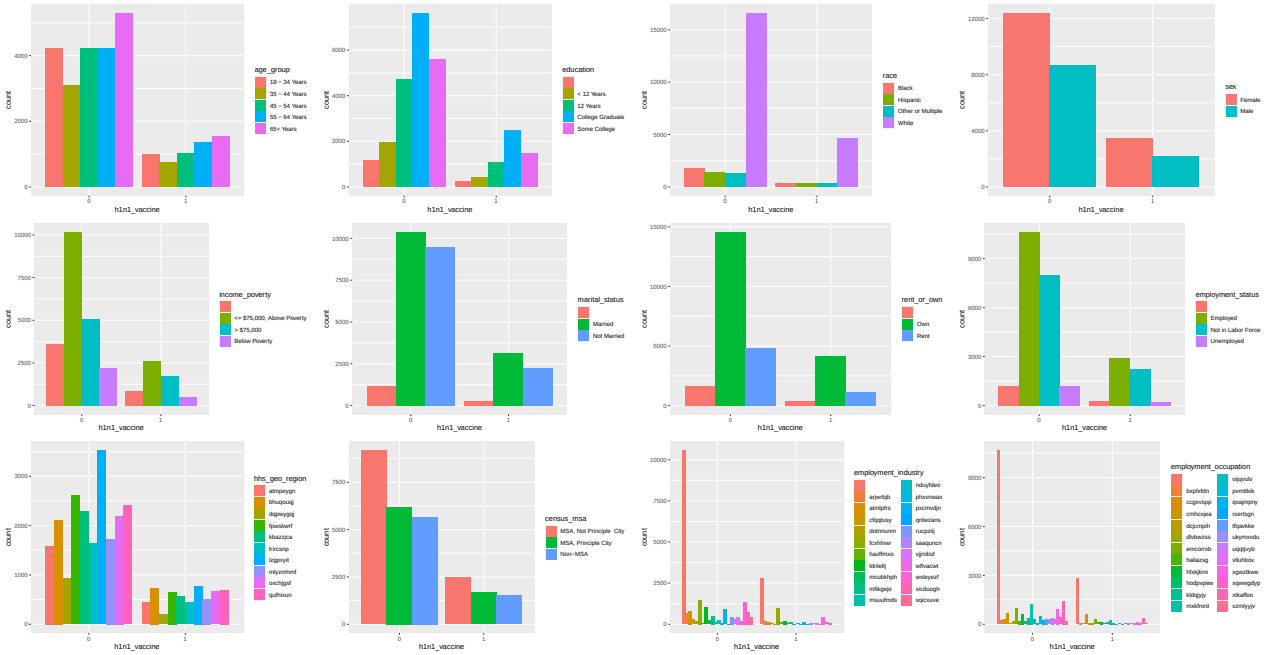
```
##
## [[8]]
```

```
##
## [[9]]
```

```
##
## [[10]]
```

```
##
## [[11]]
```

```
##
## [[12]]
```




```
lapply(categorical,
       function(x) ggplot(train, aes(seasonal_vaccine, ..count..))
       + geom_bar(aes_string(fill=x), position="dodge"))
```

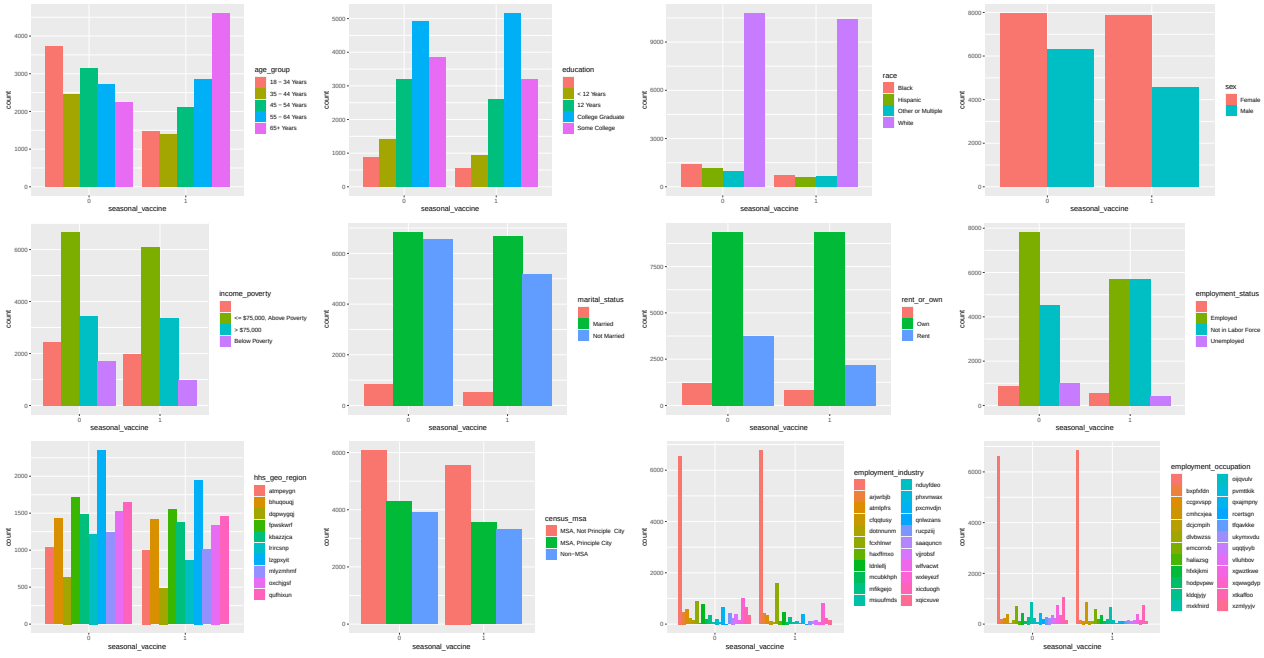
```
## [[1]]
```

```
##
## [[2]]
```

```
##
## [[3]]
```

```
##
## [[4]]
```

```
##
## [[5]]

##
## [[6]]

##
## [[7]]

##
## [[8]]

##
## [[9]]

##
## [[10]]

##
## [[11]]

##
## [[12]]
```




```
binary = c(binary, 'sex', 'marital_status', 'rent_or_own')

train = train %>%
  mutate(sex = if_else(sex == "Male", 1,
                       if_else(sex == "Female", 0,
                               median(as.numeric(sex)-1, na.rm=T)))) %>%
```

```
  mutate(marital_status = if_else(marital_status == "Married", 1,
                                  if_else(marital_status == "Not Married", 0,
                                          median(as.numeric(marital_status)-1, na.rm=T)))) %>%
  mutate(rent_or_own = if_else(rent_or_own == "Own", 1,
                               if_else(rent_or_own == "Rent", 0,
                                       median(as.numeric(rent_or_own)-1, na.rm=T))))


test = test %>%
  mutate(sex = if_else(sex == "Female", 1,
                       if_else(sex == "Male", 0,
                               median(as.numeric(sex)-1, na.rm=T)))) %>%
  mutate(marital_status = if_else(marital_status == "Married", 1,
                                  if_else(marital_status == "Not Married", 0,
                                          median(as.numeric(marital_status)-1, na.rm=T)))) %>%
  mutate(rent_or_own = if_else(rent_or_own == "Own", 1,
                               if_else(rent_or_own == "Rent", 0,
                                       median(as.numeric(rent_or_own)-1, na.rm=T))))
```

```
# Variabel target akan dikembalikan ke dalam numeric.
train$h1n1_vaccine = as.numeric(train$h1n1_vaccine)-1
train$seasonal_vaccine = as.numeric(train$seasonal_vaccine)-1
```

```
# Uji Korelasi Pearson untuk variabel dependen numerik.
cors = NULL
cors = matrix(nrow = length(numerical), ncol = 2)
rownames(cors) = numerical
colnames(cors) = c("h1n1_vaccine", "seasonal_vaccine")
for(var in numerical){
  cors[paste(var), 1] = cor(train$h1n1_vaccine, train[, paste(var)])
  cors[paste(var), 2] = cor(train$seasonal_vaccine, train[, paste(var)])
}
cors
```

```
##                              h1n1_vaccine seasonal_vaccine
## h1n1_concern                  0.121573813       0.15448822
## h1n1_knowledge                0.117771235       0.11977870
## opinion_h1n1_vacc_effective   0.267351725       0.20318688
## opinion_h1n1_risk             0.320580080       0.21565019
## opinion_h1n1_sick_from_vacc   0.074580215       0.02796437
## opinion_seas_vacc_effective   0.177798564       0.35886879
## opinion_seas_risk             0.255874248       0.38691570
## opinion_seas_sick_from_vacc   0.008415054      -0.06053783
## household_adults              0.007323224      -0.06513675
## household_children           -0.002566816      -0.11167951
```

Variabel dengan nilai korelasi terhadap dependen variabel yang lebih kecil dari 0.05

```
removeVar_a = c('opinion_seas_sick_from_vacc','household_adults','household_children','employment_occupa
removeVar_b = c('opinion_h1n1_sick_from_vacc','employment_occupation')
```

```
train_a = train[, !names(train)%in%c("seasonal_vaccine", removeVar_a)]
test_a = test[, !names(test)%in%c("seasonal_vaccine", removeVar_a)]
train_b = train[, !names(train)%in%c("h1n1_vaccine", removeVar_b)]
test_b = test[, !names(test)%in%c("h1n1_vaccine", removeVar_b)]
```

4 Train-Test Split pada Train Set

Akan dilakukan pemisahan Train-Test pada data training yang diberikan untuk melakukan validasi lokal. Rasio Train-Test sebesar 80:20 dengan menerapkan stratified random sampling terhadap independent variable. Proses ini dilakukan untuk dua kasus, yaitu untuk pemodelan model untuk (a) h1n1_vaccine, dan (b) seasonal vaccine.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

```
## Loading required package: lattice
```

```
set.seed(1)
train_idx_a <- createDataPartition(y=train_a$h1n1_vaccine, p=0.8, list=F)
train_idx_b <- createDataPartition(y=train_b$seasonal_vaccine, p=0.8, list=F)

train_a <- train_a[train_idx_a,]
testloc_a <- train_a[-train_idx_a,]

train_b <- train_b[train_idx_b,]
testloc_b <- train_b[-train_idx_b,]
```

5 Model

Akan dibuat fungsi untuk mengevaluasi model secara lokal dengan metrik ROC-AUC. AUC yang akan diambil adalah rataan AUC dari 2 variabel target.

```
library(ROCR) # Kalkulasi AUC
```

```
## Warning: package 'ROCR' was built under R version 4.1.2
```

```
roc_auc <- function(pred_a, pred_b){
  obs_a = testloc_a$h1n1_vaccine
  obs_b = testloc_b$seasonal_vaccine

  ROCPred_a <- prediction(as.numeric(pred_a), as.numeric(obs_a))
  auc_a <- performance(ROCPred_a, measure = "auc")
  auc_a <- auc_a@y.values[[1]]

  ROCPred_b <- prediction(as.numeric(pred_b), as.numeric(obs_b))
  auc_b <- performance(ROCPred_b, measure = "auc")
  auc_b <- auc_b@y.values[[1]]
  return(mean(c(auc_a, auc_b)))
}
```

5.1 Logistic Regression

```
library(car) # VIF
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

5.1.1 Model A

```
logr_a = glm(h1n1_vaccine ~ ., family = "binomial", data = train_a)
summary(logr_a)
```

```
##
## Call:
## glm(formula = h1n1_vaccine ~ ., family = "binomial", data = train_a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5675  -0.5450  -0.3106  -0.1291   3.2397
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -6.004779   0.213923 -28.070  < 2e-16
## h1n1_concern                         0.027444   0.028548   0.961 0.336386
## h1n1_knowledge                       0.209695   0.037987   5.520 3.39e-08
## behavioral_antiviral_meds            0.163857   0.092618   1.769 0.076865
## behavioral_avoidance                -0.085794   0.053374  -1.607 0.107966
## behavioral_face_mask                 0.156113   0.077695   2.009 0.044505
## behavioral_wash_hands                0.083868   0.067330   1.246 0.212904
## behavioral_large_gatherings         -0.095538   0.054530  -1.752 0.079772
## behavioral_outside_home             -0.004172   0.055356  -0.075 0.939919
## behavioral_touch_face               -0.014926   0.051767  -0.288 0.773095
## doctor_recc_h1n1                     1.991753   0.061594  32.337  < 2e-16
## doctor_recc_seasonal                -0.552703   0.060310  -9.164  < 2e-16
## chronic_med_condition                0.137515   0.048101   2.859 0.004252
## child_under_6_months                 0.284863   0.072976   3.904 9.48e-05
## health_worker                        0.787976   0.085473   9.219  < 2e-16
## health_insurance                     0.816925   0.024978  32.705  < 2e-16
## opinion_h1n1_vacc_effective          0.572592   0.028941  19.785  < 2e-16
```

```
## opinion_h1n1_risk                          0.423299   0.020238  20.916  < 2e-16
## opinion_h1n1_sick_from_vacc               -0.033148   0.017081  -1.941 0.052298
## opinion_seas_vacc_effective                0.131043   0.026643   4.918 8.72e-07
## opinion_seas_risk                          0.123945   0.019330   6.412 1.44e-10
## age_group35 - 44 Years                    -0.045370   0.077116  -0.588 0.556309
## age_group45 - 54 Years                     0.020733   0.072241   0.287 0.774117
## age_group55 - 64 Years                     0.330680   0.070536   4.688 2.76e-06
## age_group65+ Years                         0.466621   0.074514   6.262 3.80e-10
## education< 12 Years                       -0.971521   0.199904  -4.860 1.17e-06
## education12 Years                         -0.860036   0.190731  -4.509 6.51e-06
## educationCollege Graduate                 -0.679661   0.189421  -3.588 0.000333
## educationSome College                     -0.824975   0.190188  -4.338 1.44e-05
## raceHispanic                               0.307674   0.118264   2.602 0.009280
## raceOther or Multiple                      0.373805   0.119908   3.117 0.001824
## raceWhite                                  0.232920   0.089862   2.592 0.009542
## sex                                        0.178809   0.046101   3.879 0.000105
## income_poverty<= $75,000, Above Poverty   -0.186180   0.071304  -2.611 0.009026
## income_poverty> $75,000                   -0.113975   0.080997  -1.407 0.159381
## income_povertyBelow Poverty               -0.149505   0.098312  -1.521 0.128333
## marital_status                             0.102089   0.047331   2.157 0.031012
## rent_or_own                                0.085232   0.058805   1.449 0.147229
## employment_statusEmployed                 -0.719146   0.353796  -2.033 0.042087
## employment_statusNot in Labor Force       -0.306991   0.189835  -1.617 0.105846
## employment_statusUnemployed               -0.331171   0.210370  -1.574 0.115435
## hhs_geo_regionbhuqouqj                    -0.052602   0.096695  -0.544 0.586445
## hhs_geo_regiondqpwygqj                    -0.295235   0.133012  -2.220 0.026445
## hhs_geo_regionfpwskwrf                    -0.148286   0.096568  -1.536 0.124645
## hhs_geo_regionkbazzjca                    -0.168381   0.098725  -1.706 0.088090
## hhs_geo_regionlrircsnp                     0.016270   0.105049   0.155 0.876915
## hhs_geo_regionlzgpxyit                    -0.193407   0.092545  -2.090 0.036630
## hhs_geo_regionmlyzmhmf                    -0.025665   0.104361  -0.246 0.805742
## hhs_geo_regionoxchjgsf                     0.147836   0.096349   1.534 0.124935
## hhs_geo_regionqufhixun                    -0.010659   0.095848  -0.111 0.911452
## census_msaMSA, Principle City              0.076999   0.051204   1.504 0.132646
## census_msaNon-MSA                          0.068867   0.053458   1.288 0.197658
## employment_industryarjwrbjb                0.771225   0.322169   2.394 0.016672
## employment_industryatmlpfrs                0.076065   0.328524   0.232 0.816898
## employment_industrycfqqtusy                0.084293   0.367510   0.229 0.818587
## employment_industrydotnnunm                0.011060   0.420064   0.026 0.978995
## employment_industryfcxhlnwr                0.596728   0.315903   1.889 0.058897
## employment_industryhaxffmxo                2.899455   0.384395   7.543 4.60e-14
## employment_industryldnlellj                0.303121   0.319891   0.948 0.343345
## employment_industrymcubkhph                0.229288   0.384735   0.596 0.551199
## employment_industrymfikgejo                0.217475   0.338398   0.643 0.520444
## employment_industrymsuufmds                0.344841   0.461511   0.747 0.454942
## employment_industrynduyfdeo                0.594519   0.367283   1.619 0.105513
## employment_industryphxvnwax                0.192177   0.508493   0.378 0.705479
## employment_industrypxcmvdjn               -0.035993   0.327307  -0.110 0.912436
## employment_industryqnlwzans               -0.196826   1.228736  -0.160 0.872735
## employment_industryrucpziij               -0.046172   0.349398  -0.132 0.894868
## employment_industrysaaquncn                0.609243   0.356855   1.707 0.087774
## employment_industryvjjrobsf                0.045027   0.347709   0.129 0.896965
## employment_industrywlfvacwt                0.074475   0.402536   0.185 0.853218
## employment_industrywxleyezf                0.438873   0.313056   1.402 0.160945
```

```
## employment_industryxicduogh              0.293852   0.332308   0.884 0.376547
## employment_industryxqicxuve              0.268077   0.347427   0.772 0.440348
##
## (Intercept)                             ***
## h1n1_concern
## h1n1_knowledge                          ***
## behavioral_antiviral_meds               .
## behavioral_avoidance
## behavioral_face_mask                    *
## behavioral_wash_hands
## behavioral_large_gatherings             .
## behavioral_outside_home
## behavioral_touch_face
## doctor_recc_h1n1                        ***
## doctor_recc_seasonal                    ***
## chronic_med_condition                   **
## child_under_6_months                    ***
## health_worker                           ***
## health_insurance                        ***
## opinion_h1n1_vacc_effective             ***
## opinion_h1n1_risk                       ***
## opinion_h1n1_sick_from_vacc             .
## opinion_seas_vacc_effective             ***
## opinion_seas_risk                       ***
## age_group35 - 44 Years
## age_group45 - 54 Years
## age_group55 - 64 Years                  ***
## age_group65+ Years                      ***
## education< 12 Years                     ***
## education12 Years                       ***
## educationCollege Graduate               ***
## educationSome College                   ***
## raceHispanic                            **
## raceOther or Multiple                   **
## raceWhite                               **
## sex                                     ***
## income_poverty<= $75,000, Above Poverty **
## income_poverty> $75,000
## income_povertyBelow Poverty
## marital_status                          *
## rent_or_own
## employment_statusEmployed               *
## employment_statusNot in Labor Force
## employment_statusUnemployed
## hhs_geo_regionbhuqouqj
## hhs_geo_regiondqpwygqj                  *
## hhs_geo_regionfpwskwrf
## hhs_geo_regionkbazzjca                  .
## hhs_geo_regionlrircsnp
## hhs_geo_regionlzgpxyit                  *
## hhs_geo_regionmlyzmhmf
## hhs_geo_regionoxchjgsf
## hhs_geo_regionqufhixun
## census_msaMSA, Principle City
```

```
## census_msaNon-MSA
## employment_industryarjwrbjb             *
## employment_industryatmlpfrs
## employment_industrycfqqtusy
## employment_industrydotnnunm
## employment_industryfcxhlnwr             .
## employment_industryhaxffmxo             ***
## employment_industryldnlellj
## employment_industrymcubkhph
## employment_industrymfikgejo
## employment_industrymsuufmds
## employment_industrynduyfdeo
## employment_industryphxvnwax
## employment_industrypxcmvdjn
## employment_industryqnlwzans
## employment_industryrucpziij
## employment_industrysaaquncn             .
## employment_industryvjjrobsf
## employment_industrywlfvacwt
## employment_industrywxleyezf
## employment_industryxicduogh
## employment_industryxqicxuve
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22042  on 21365  degrees of freedom
## Residual deviance: 14908  on 21293  degrees of freedom
## AIC: 15054
##
## Number of Fisher Scoring iterations: 6
```

```
vif(logr_a)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## h1n1_concern                1.450824  1        1.204502
## h1n1_knowledge              1.231872  1        1.109897
## behavioral_antiviral_meds   1.077509  1        1.038031
## behavioral_avoidance        1.227406  1        1.107884
## behavioral_face_mask        1.098119  1        1.047912
## behavioral_wash_hands       1.265152  1        1.124790
## behavioral_large_gatherings 1.597509  1        1.263926
## behavioral_outside_home     1.606152  1        1.267341
## behavioral_touch_face       1.260591  1        1.122760
## doctor_recc_h1n1            1.939900  1        1.392803
## doctor_recc_seasonal        1.979788  1        1.407049
## chronic_med_condition       1.147721  1        1.071317
## child_under_6_months        1.042915  1        1.021232
## health_worker               2.150102  1        1.466323
## health_insurance            1.230844  1        1.109434
## opinion_h1n1_vacc_effective 1.206988  1        1.098630
## opinion_h1n1_risk           1.660996  1        1.288796
## opinion_h1n1_sick_from_vacc 1.290553  1        1.136025
```

```
## opinion_seas_vacc_effective   1.265447  1        1.124921
## opinion_seas_risk             1.590457  1        1.261133
## age_group                     1.830828  4        1.078527
## education                     6.459110  4        1.262617
## race                          1.415169  3        1.059582
## sex                           1.172889  1        1.083000
## income_poverty                2.388380  3        1.156158
## marital_status                1.256662  1        1.121009
## rent_or_own                   1.308874  1        1.144060
## employment_status           234.033665  3        2.482445
## hhs_geo_region                1.341713  9        1.016464
## census_msa                    1.182722  2        1.042847
## employment_industry         147.851405 21        1.126322
```

```
logrPredict_a = predict(logr_a, type="response", newdata=testloc_a)
logrPredict_a = ifelse(logrPredict_a > 0.5, 1, 0)
```

5.1.2 Model B

```
logr_b = glm(seasonal_vaccine ~ ., family = "binomial", data = train_b)
summary(logr_b)
```

```
##
## Call:
## glm(formula = seasonal_vaccine ~ ., family = "binomial", data = train_b)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9957  -0.7318  -0.2512   0.7296   3.1908
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -5.2440977  0.1734293 -30.238  < 2e-16
## h1n1_concern                      0.0402325  0.0235422   1.709 0.087459
## h1n1_knowledge                    0.1815619  0.0314564   5.772 7.84e-09
## behavioral_antiviral_meds         0.0494852  0.0831327   0.595 0.551673
## behavioral_avoidance             -0.0385614  0.0442609  -0.871 0.383629
## behavioral_face_mask              0.0546003  0.0723634   0.755 0.450531
## behavioral_wash_hands             0.1156510  0.0533694   2.167 0.030236
## behavioral_large_gatherings      -0.0105368  0.0464205  -0.227 0.820434
## behavioral_outside_home          -0.0300180  0.0472283  -0.636 0.525042
## behavioral_touch_face             0.2374198  0.0427225   5.557 2.74e-08
## doctor_recc_h1n1                 -0.3370353  0.0584217  -5.769 7.97e-09
## doctor_recc_seasonal              1.4779701  0.0508556  29.062  < 2e-16
## chronic_med_condition             0.2194258  0.0416934   5.263 1.42e-07
## child_under_6_months              0.1284679  0.0669393   1.919 0.054963
## health_worker                     0.6791599  0.0781559   8.690  < 2e-16
## health_insurance                  0.2314642  0.0190596  12.144  < 2e-16
## opinion_h1n1_vacc_effective       0.0012017  0.0211387   0.057 0.954666
## opinion_h1n1_risk                 0.0395285  0.0177331   2.229 0.025809
## opinion_seas_vacc_effective       0.5768869  0.0220336  26.182  < 2e-16
```

```
## opinion_seas_risk                          0.5340483  0.0164953  32.376  < 2e-16
## opinion_seas_sick_from_vacc               -0.2267445  0.0149090 -15.209  < 2e-16
## age_group35 - 44 Years                     0.2723534  0.0649812   4.191 2.77e-05
## age_group45 - 54 Years                     0.4721424  0.0608479   7.759 8.53e-15
## age_group55 - 64 Years                     0.7809631  0.0631667  12.364  < 2e-16
## age_group65+ Years                         1.6074918  0.0677045  23.743  < 2e-16
## education< 12 Years                       -0.4034131  0.1615342  -2.497 0.012511
## education12 Years                         -0.2136520  0.1537137  -1.390 0.164549
## educationCollege Graduate                 -0.0111342  0.1528382  -0.073 0.941926
## educationSome College                     -0.1603050  0.1537708  -1.042 0.297183
## raceHispanic                               0.2472514  0.0982997   2.515 0.011894
## raceOther or Multiple                      0.3786053  0.0988063   3.832 0.000127
## raceWhite                                  0.3440799  0.0714273   4.817 1.46e-06
## sex                                        0.0289727  0.0384024   0.754 0.450579
## income_poverty<= $75,000, Above Poverty   -0.0924693  0.0589676  -1.568 0.116849
## income_poverty> $75,000                    0.0214670  0.0677129   0.317 0.751221
## income_povertyBelow Poverty               -0.1807094  0.0822100  -2.198 0.027939
## marital_status                             0.1383936  0.0436607   3.170 0.001526
## rent_or_own                                0.1682395  0.0496729   3.387 0.000707
## employment_statusEmployed                 -0.6180260  0.2661767  -2.322 0.020240
## employment_statusNot in Labor Force       -0.2055804  0.1513744  -1.358 0.174434
## employment_statusUnemployed               -0.5324007  0.1687100  -3.156 0.001601
## hhs_geo_regionbhuqouqj                    -0.1766303  0.0835675  -2.114 0.034547
## hhs_geo_regiondqpwygqj                    -0.2420986  0.1085454  -2.230 0.025722
## hhs_geo_regionfpwskwrf                    -0.0539715  0.0818743  -0.659 0.509768
## hhs_geo_regionkbazzjca                     0.0001236  0.0833757   0.001 0.998817
## hhs_geo_regionlrircsnp                    -0.2346826  0.0893955  -2.625 0.008659
## hhs_geo_regionlzgpxyit                    -0.1547502  0.0779120  -1.986 0.047009
## hhs_geo_regionmlyzmhmf                    -0.1347782  0.0897337  -1.502 0.133103
## hhs_geo_regionoxchjgsf                     0.0339792  0.0832421   0.408 0.683129
## hhs_geo_regionqufhixun                    -0.1164690  0.0816373  -1.427 0.153677
## census_msaMSA, Principle City              0.0224147  0.0430414   0.521 0.602526
## census_msaNon-MSA                         -0.1459363  0.0450277  -3.241 0.001191
## household_adults                          -0.0260828  0.0273906  -0.952 0.340969
## household_children                        -0.0429035  0.0233240  -1.839 0.065848
## employment_industryarjwrbjb                0.6863042  0.2456756   2.794 0.005213
## employment_industryatmlpfrs                0.4862861  0.2449712   1.985 0.047136
## employment_industrycfqqtusy                0.1928813  0.2771025   0.696 0.486388
## employment_industrydotnnunm                0.2408699  0.2950828   0.816 0.414341
## employment_industryfcxhlnwr                0.6298727  0.2416424   2.607 0.009144
## employment_industryhaxffmxo                3.2603480  0.3540502   9.209  < 2e-16
## employment_industryldnlellj                0.2047024  0.2415015   0.848 0.396647
## employment_industrymcubkhph               -0.1802999  0.2947161  -0.612 0.540687
## employment_industrymfikgejo                0.5635344  0.2525315   2.232 0.025645
## employment_industrymsuufmds                0.9516349  0.3359335   2.833 0.004614
## employment_industrynduyfdeo               -0.0357549  0.2904699  -0.123 0.902033
## employment_industryphxvnwax                0.1660173  0.3829564   0.434 0.664641
## employment_industrypxcmvdjn                0.1481469  0.2439007   0.607 0.543581
## employment_industryqnlwzans               -0.4938720  0.7846238  -0.629 0.529062
## employment_industryrucpziij               -0.2438765  0.2702510  -0.902 0.366840
## employment_industrysaaquncn               -0.1965239  0.2842127  -0.691 0.489272
## employment_industryvjjrobsf               -0.1414442  0.2617258  -0.540 0.588901
## employment_industrywlfvacwt                0.2559402  0.2994353   0.855 0.392693
## employment_industrywxleyezf                0.2543482  0.2369272   1.074 0.283034
```

```
## employment_industryxicduogh              0.0031606  0.2525059   0.013 0.990013
## employment_industryxqicxuve             -0.0567326  0.2613060  -0.217 0.828121
##
## (Intercept)                              ***
## h1n1_concern                             .
## h1n1_knowledge                           ***
## behavioral_antiviral_meds
## behavioral_avoidance
## behavioral_face_mask
## behavioral_wash_hands                    *
## behavioral_large_gatherings
## behavioral_outside_home
## behavioral_touch_face                    ***
## doctor_recc_h1n1                         ***
## doctor_recc_seasonal                     ***
## chronic_med_condition                    ***
## child_under_6_months                     .
## health_worker                            ***
## health_insurance                         ***
## opinion_h1n1_vacc_effective
## opinion_h1n1_risk                        *
## opinion_seas_vacc_effective              ***
## opinion_seas_risk                        ***
## opinion_seas_sick_from_vacc              ***
## age_group35 - 44 Years                   ***
## age_group45 - 54 Years                   ***
## age_group55 - 64 Years                   ***
## age_group65+ Years                       ***
## education< 12 Years                      *
## education12 Years
## educationCollege Graduate
## educationSome College
## raceHispanic                             *
## raceOther or Multiple                    ***
## raceWhite                                ***
## sex
## income_poverty<= $75,000, Above Poverty
## income_poverty> $75,000
## income_povertyBelow Poverty              *
## marital_status                           **
## rent_or_own                              ***
## employment_statusEmployed                *
## employment_statusNot in Labor Force
## employment_statusUnemployed              **
## hhs_geo_regionbhuqouqj                   *
## hhs_geo_regiondqpwygqj                   *
## hhs_geo_regionfpwskwrf
## hhs_geo_regionkbazzjca
## hhs_geo_regionlrircsnp                   **
## hhs_geo_regionlzgpxyit                   *
## hhs_geo_regionmlyzmhmf
## hhs_geo_regionoxchjgsf
## hhs_geo_regionqufhixun
## census_msaMSA, Principle City
```

```
## census_msaNon-MSA                        **
## household_adults
## household_children                       .
## employment_industryarjwrbjb              **
## employment_industryatmlpfrs              *
## employment_industrycfqqtusy
## employment_industrydotnnunm
## employment_industryfcxhlnwr              **
## employment_industryhaxffmxo              ***
## employment_industryldnlellj
## employment_industrymcubkhph
## employment_industrymfikgejo              *
## employment_industrymsuufmds              **
## employment_industrynduyfdeo
## employment_industryphxvnwax
## employment_industrypxcmvdjn
## employment_industryqnlwzans
## employment_industryrucpziij
## employment_industrysaaquncn
## employment_industryvjjrobsf
## employment_industrywlfvacwt
## employment_industrywxleyezf
## employment_industryxicduogh
## employment_industryxqicxuve
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29519  on 21365  degrees of freedom
## Residual deviance: 20015  on 21291  degrees of freedom
## AIC: 20165
##
## Number of Fisher Scoring iterations: 5
```

```
vif(logr_b)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## h1n1_concern                1.417414  1        1.190552
## h1n1_knowledge              1.221455  1        1.105195
## behavioral_antiviral_meds   1.078737  1        1.038623
## behavioral_avoidance        1.242863  1        1.114838
## behavioral_face_mask        1.096282  1        1.047035
## behavioral_wash_hands       1.262023  1        1.123398
## behavioral_large_gatherings 1.621916  1        1.273545
## behavioral_outside_home     1.631457  1        1.277285
## behavioral_touch_face       1.275988  1        1.129596
## doctor_recc_h1n1            1.727842  1        1.314474
## doctor_recc_seasonal        1.669279  1        1.292006
## chronic_med_condition       1.107420  1        1.052340
## child_under_6_months        1.045242  1        1.022371
## health_worker               1.868905  1        1.367079
## health_insurance            1.122069  1        1.059278
## opinion_h1n1_vacc_effective 1.254070  1        1.119853
```

```
## opinion_h1n1_risk             1.614238  1        1.270527
## opinion_seas_vacc_effective   1.216644  1        1.103016
## opinion_seas_risk             1.562792  1        1.250117
## opinion_seas_sick_from_vacc   1.213587  1        1.101629
## age_group                     2.409929  4        1.116222
## education                     6.553327  4        1.264904
## race                          1.411360  3        1.059106
## sex                           1.157626  1        1.075930
## income_poverty                2.460736  3        1.161923
## marital_status                1.524888  1        1.234864
## rent_or_own                   1.318307  1        1.148175
## employment_status           189.434456  3        2.396495
## hhs_geo_region                1.342475  9        1.016497
## census_msa                    1.193038  2        1.045114
## household_adults              1.347458  1        1.160801
## household_children            1.499494  1        1.224538
## employment_industry         103.674695 21        1.116843
```

```
logrPredict_b = predict(logr_b, type="response", newdata=testloc_b)
logrPredict_b = ifelse(logrPredict_b > 0.5, 1, 0)
```

5.1.3 Prediction

```
# Evaluasi AUC secara lokal
(logrAUC = roc_auc(logrPredict_a, logrPredict_b))
```

```
## [1] 0.7481868
```

```
predict_a = predict(logr_a, type="response", newdata=test)
predict_b = predict(logr_b, type="response", newdata=test)
```

```
submission <- cbind("respondent_id"=as.numeric(rownames(test)),
                    "h1n1_vaccine"=predict_a,
                    "seasonal_vaccine"=predict_b)
head(submission)
```

```
##       respondent_id h1n1_vaccine seasonal_vaccine
## 26707         26707   0.14003927       0.31832347
## 26708         26708   0.02099115       0.04992814
## 26709         26709   0.33996869       0.49861052
## 26710         26710   0.60949851       0.89644628
## 26711         26711   0.36861935       0.54964695
## 26712         26712   0.69853584       0.93948178
```

```
# write.csv(submission,"submission_logr.csv", row.names = FALSE)
```

5.2 Naive Bayes

```r
library(e1071) # Pemodelan Naive Bayes
```

```
## Warning: package 'e1071' was built under R version 4.1.3
```

```r
options = trainControl(method="repeatedCV", number=10) # 10-fold cross validation utk akurasi
```

5.2.1 Model A

```r
nb_a = naiveBayes(train_a, train_a$h1n1_vaccine, laplace=1, trControl=options, tuneLength=7)
nb_a
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train_a, y = train_a$h1n1_vaccine, laplace = 1,
##     trControl = options, tuneLength = 7)
##
## A-priori probabilities:
## train_a$h1n1_vaccine
##         0         1
## 0.7885893 0.2114107
##
## Conditional probabilities:
##                      h1n1_vaccine
## train_a$h1n1_vaccine [,1] [,2]
##                    0    0    0
##                    1    1    0
##
##                      h1n1_concern
## train_a$h1n1_vaccine     [,1]      [,2]
##                    0 1.559202 0.9110875
##                    1 1.837281 0.8706668
##
##                      h1n1_knowledge
## train_a$h1n1_vaccine     [,1]      [,2]
##                    0 1.221081 0.6158264
##                    1 1.403144 0.6013116
##
##                      behavioral_antiviral_meds
## train_a$h1n1_vaccine       [,1]      [,2]
##                    0 0.04314796 0.2031961
##                    1 0.06752269 0.2509528
##
##                      behavioral_avoidance
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.7178468 0.4500609
##                    1 0.7662165 0.4232829
##
##                      behavioral_face_mask
```

```
## train_a$h1n1_vaccine       [,1]      [,2]
##                    0 0.05958811 0.2367291
##                    1 0.10449413 0.3059344
##
##                     behavioral_wash_hands
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.8114428 0.3911681
##                    1 0.8808944 0.3239486
##
##                     behavioral_large_gatherings
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.3509407 0.4772786
##                    1 0.3816692 0.4858499
##
##                     behavioral_outside_home
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.3285061 0.4696839
##                    1 0.3595307 0.4799160
##
##                     behavioral_touch_face
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.6633035 0.4725942
##                    1 0.7427496 0.4371669
##
##                     doctor_recc_h1n1
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.1206600 0.3257414
##                    1 0.5096303 0.4999626
##
##                     doctor_recc_seasonal
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.2513502 0.4338023
##                    1 0.4965685 0.5000436
##
##                     chronic_med_condition
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.2498071 0.4329141
##                    1 0.3593093 0.4798511
##
##                     child_under_6_months
## train_a$h1n1_vaccine       [,1]      [,2]
##                    0 0.07086474 0.2566064
##                    1 0.11334957 0.3170547
##
##                     health_worker
## train_a$h1n1_vaccine       [,1]      [,2]
##                    0 0.08071696 0.2724080
##                    1 0.20943104 0.4069476
##
##                     health_insurance
## train_a$h1n1_vaccine       [,1]      [,2]
##                    0 -0.1056443 0.9582312
##                    1  0.4571618 0.8626900
##
```

```
##                      opinion_h1n1_vacc_effective
## train_a$h1n1_vaccine     [,1]      [,2]
##                    0 3.710012 1.0087792
##                    1 4.372814 0.7599621
##
##                      opinion_h1n1_risk
## train_a$h1n1_vaccine     [,1]     [,2]
##                    0 2.122915 1.171501
##                    1 3.129511 1.339695
##
##                      opinion_h1n1_sick_from_vacc
## train_a$h1n1_vaccine     [,1]     [,2]
##                    0 2.306250 1.325246
##                    1 2.554793 1.447004
##
##                      opinion_seas_vacc_effective
## train_a$h1n1_vaccine     [,1]      [,2]
##                    0 3.922251 1.1152739
##                    1 4.401151 0.8233089
##
##                      opinion_seas_risk
## train_a$h1n1_vaccine     [,1]     [,2]
##                    0 2.519853 1.333280
##                    1 3.384104 1.312083
##
##                      age_group
## train_a$h1n1_vaccine 18 - 34 Years 35 - 44 Years 45 - 54 Years 55 - 64 Years
##                    0     0.2013172     0.1454254     0.2023852     0.2009612
##                    1     0.1722689     0.1329058     0.1806723     0.2408226
##                      age_group
## train_a$h1n1_vaccine 65+ Years
##                    0 0.2499110
##                    1 0.2733304
##
##                      education
## train_a$h1n1_vaccine            < 12 Years   12 Years College Graduate
##                    0 0.05339979 0.09303429 0.22647443       0.36299988
##                    1 0.04533392 0.07386112 0.18620080       0.43675365
##                      education
## train_a$h1n1_vaccine Some College
##                    0   0.26409161
##                    1   0.25785051
##
##                      race
## train_a$h1n1_vaccine      Black   Hispanic Other or Multiple      White
##                    0 0.08514804 0.06675369        0.05933662 0.78876164
##                    1 0.05861535 0.06525105        0.06193320 0.81420040
##
##                      sex
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.4127841 0.4923492
##                    1 0.3838831 0.4863838
##
##                      income_poverty
```

```
## train_a$h1n1_vaccine           <= $75,000, Above Poverty  > $75,000
##                    0 0.17112680                0.48199134 0.24209340
##                    1 0.14554302                0.46073877 0.30369387
##                     income_poverty
## train_a$h1n1_vaccine Below Poverty
##                    0    0.10478846
##                    1    0.09002433
##
##                     marital_status
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.5467980 0.4978199
##                    1 0.5988488 0.4901858
##
##                     rent_or_own
## train_a$h1n1_vaccine      [,1]      [,2]
##                    0 0.7706689 0.4204151
##                    1 0.8029666 0.3978018
##
##                     employment_status
## train_a$h1n1_vaccine              Employed Not in Labor Force Unemployed
##                    0 0.05619178 0.50394589         0.38183113 0.05803121
##                    1 0.04578633 0.52023889         0.39172749 0.04224729
##
##                     hhs_geo_region
## train_a$h1n1_vaccine   atmpeygn   bhuqouqj   dqpwygqj   fpwskwrf   kbazzjca
##                    0 0.07663562 0.10119224 0.04365621 0.12272377 0.11264013
##                    1 0.08084824 0.12878286 0.03313453 0.11508725 0.09962448
##                     hhs_geo_region
## train_a$h1n1_vaccine   lrircsnp   lzgpxyit   mlyzmhmf   oxchjgsf   qufhixun
##                    0 0.07823714 0.16726971 0.08144018 0.10320897 0.11299603
##                    1 0.07797658 0.13629335 0.09078860 0.11928429 0.11817981
##
##                     census_msa
## train_a$h1n1_vaccine MSA, Not Principle  City MSA, Principle City   Non-MSA
##                    0                0.4328863           0.2946831 0.2724306
##                    1                0.4342920           0.2951327 0.2705752
##
##                     employment_industry
## train_a$h1n1_vaccine                  arjwrbjb     atmlpfrs     cfqqtusy
##                    0 0.5031118487 0.0312963073 0.0380534645 0.0136328611
##                    1 0.4811632518 0.0376734964 0.0231328486 0.0085922009
##                     employment_industry
## train_a$h1n1_vaccine     dotnnunm     fcxhlnwr     haxffmxo     ldnlellj
##                    0 0.0088317231 0.0696461383 0.0029043922 0.0493746666
##                    1 0.0041859440 0.1747080855 0.0158625248 0.0383344349
##                     employment_industry
## train_a$h1n1_vaccine     mcubkhph     mfikgejo     msuufmds     nduyfdeo
##                    0 0.0109655622 0.0245984233 0.0048604114 0.0113212021
##                    1 0.0072703239 0.0171844019 0.0030843798 0.0092531395
##                     employment_industry
## train_a$h1n1_vaccine     phxvnwax     pxcmvdjn     qnlwzans     rucpziij
##                    0 0.0036156719 0.0422618695 0.0005334598 0.0209827515
##                    1 0.0022031284 0.0244547257 0.0004406257 0.0123375193
##                     employment_industry
```

```
## train_a$h1n1_vaccine     saaquncn     vjjrobsf     wlfvacwt     wxleyezf
##                    0 0.0127437615 0.0221089443 0.0089502697 0.0650820935
##                    1 0.0114562679 0.0127781450 0.0050671954 0.0775501212
##                     employment_industry
## train_a$h1n1_vaccine     xicduogh     xqicxuve
##                    0 0.0356825322 0.0194416454
##                    1 0.0193875303 0.0138797092
```

```
nbPredict_a = predict(nb_a, type="class", newdata=testloc_a)
```

5.2.2 Model B

```
nb_b = naiveBayes(train_b, train_b$seasonal_vaccine, laplace=1, trControl=options, tuneLength=7)
nb_b
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train_b, y = train_b$seasonal_vaccine,
##     laplace = 1, trControl = options, tuneLength = 7)
##
## A-priori probabilities:
## train_b$seasonal_vaccine
##         0         1
## 0.5343536 0.4656464
##
## Conditional probabilities:
##                          seasonal_vaccine
## train_b$seasonal_vaccine [,1] [,2]
##                        0    0    0
##                        1    1    0
##
##                          h1n1_concern
## train_b$seasonal_vaccine     [,1]      [,2]
##                        0 1.484365 0.9199503
##                        1 1.771133 0.8666560
##
##                          h1n1_knowledge
## train_b$seasonal_vaccine     [,1]      [,2]
##                        0 1.190856 0.6152356
##                        1 1.342245 0.6111527
##
##                          behavioral_antiviral_meds
## train_b$seasonal_vaccine       [,1]      [,2]
##                        0 0.04817378 0.2141427
##                        1 0.05075887 0.2195159
##
##                          behavioral_avoidance
## train_b$seasonal_vaccine      [,1]      [,2]
##                        0 0.6940527 0.4608277
```

```
##                         1 0.7641974 0.4245207
##
##                          behavioral_face_mask
## train_b$seasonal_vaccine       [,1]      [,2]
##                         0 0.05588158 0.2297030
##                         1 0.08352598 0.2766895
##
##                          behavioral_wash_hands
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.7854953 0.4104963
##                         1 0.8733541 0.3325926
##
##                          behavioral_large_gatherings
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.3288079 0.4698006
##                         1 0.3932053 0.4884863
##
##                          behavioral_outside_home
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.3112902 0.4630414
##                         1 0.3643582 0.4812739
##
##                          behavioral_touch_face
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.6293247 0.4830068
##                         1 0.7407780 0.4382297
##
##                          doctor_recc_h1n1
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.1269160 0.3328934
##                         1 0.2930948 0.4552044
##
##                          doctor_recc_seasonal
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.1460103 0.3531320
##                         1 0.4818575 0.4996959
##
##                          chronic_med_condition
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.2020671 0.4015596
##                         1 0.3517942 0.4775541
##
##                          child_under_6_months
## train_b$seasonal_vaccine       [,1]      [,2]
##                         0 0.07497591 0.2633640
##                         1 0.08342547 0.2765381
##
##                          health_worker
## train_b$seasonal_vaccine       [,1]      [,2]
##                         0 0.07164754 0.2579147
##                         1 0.15096995 0.3580376
##
##                          health_insurance
## train_b$seasonal_vaccine        [,1]      [,2]
```

```
##                         0 -0.08390996 0.9484623
##                         1  0.13649613 0.9749908
##
##                          opinion_h1n1_vacc_effective
## train_b$seasonal_vaccine     [,1]      [,2]
##                         0 3.660243 1.0527887
##                         1 4.072168 0.8914881
##
##                          opinion_h1n1_risk
## train_b$seasonal_vaccine     [,1]     [,2]
##                         0 2.073662 1.177600
##                         1 2.633129 1.319068
##
##                          opinion_seas_vacc_effective
## train_b$seasonal_vaccine     [,1]      [,2]
##                         0 3.666550 1.1531908
##                         1 4.442155 0.7918691
##
##                          opinion_seas_risk
## train_b$seasonal_vaccine     [,1]     [,2]
##                         0 2.205658 1.224356
##                         1 3.275706 1.317307
##
##                          opinion_seas_sick_from_vacc
## train_b$seasonal_vaccine     [,1]     [,2]
##                         0 2.186651 1.330181
##                         1 2.030154 1.297168
##
##                          age_group
## train_b$seasonal_vaccine 18 - 34 Years 35 - 44 Years 45 - 54 Years
##                         0     0.2631763     0.1732621     0.2182630
##                         1     0.1191481     0.1128190     0.1695801
##                          age_group
## train_b$seasonal_vaccine 55 - 64 Years 65+ Years
##                         0     0.1887585 0.1565400
##                         1     0.2262407 0.3722122
##
##                          education
## train_b$seasonal_vaccine            < 12 Years   12 Years College Graduate
##                         0 0.06251094 0.09849413 0.22176501       0.34538610
##                         1 0.04390195 0.07604983 0.20946353       0.41450673
##                          education
## train_b$seasonal_vaccine Some College
##                         0   0.27184381
##                         1   0.25607796
##
##                          race
## train_b$seasonal_vaccine      Black   Hispanic Other or Multiple      White
##                         0 0.09570090 0.08151650        0.06645653 0.75632607
##                         1 0.06048428 0.04762383        0.05304933 0.83884256
##
##                          sex
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.4426732 0.4967245
```

```
##                         1 0.3676751 0.4821965
##
##                          income_poverty
## train_b$seasonal_vaccine           <= $75,000, Above Poverty > $75,000
##                         0 0.1729271                 0.4650206 0.2429735
##                         1 0.1611574                 0.4889983 0.2713755
##                          income_poverty
## train_b$seasonal_vaccine Below Poverty
##                         0     0.1190789
##                         1     0.0784688
##
##                          marital_status
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.5373566 0.4986244
##                         1 0.5858880 0.4925928
##
##                          rent_or_own
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.7370588 0.4402501
##                         1 0.8230978 0.3816051
##
##                          employment_status
## train_b$seasonal_vaccine              Employed Not in Labor Force Unemployed
##                         0 0.06461781 0.55389195         0.31118116 0.07030908
##                         1 0.04641817 0.45765096         0.46136843 0.03456244
##
##                          hhs_geo_region
## train_b$seasonal_vaccine   atmpeygn   bhuqouqj   dqpwygqj   fpwskwrf   kbazzjca
##                         0 0.07254748 0.10151396 0.04428109 0.12041656 0.10475190
##                         1 0.07922482 0.11185862 0.03865850 0.12290391 0.11205944
##                          hhs_geo_region
## train_b$seasonal_vaccine   lrircsnp   lzgpxyit   mlyzmhmf   oxchjgsf   qufhixun
##                         0 0.08707447 0.16426009 0.08514921 0.10536449 0.11464076
##                         1 0.07018777 0.15714429 0.08253841 0.10723968 0.11818456
##
##                          census_msa
## train_b$seasonal_vaccine MSA, Not Principle  City MSA, Principle City   Non-MSA
##                         0                0.4301226           0.2991243 0.2707531
##                         1                0.4463424           0.2895900 0.2640675
##
##                          household_adults
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.9286152 0.7728544
##                         1 0.8370691 0.7141012
##
##                          household_children
## train_b$seasonal_vaccine      [,1]      [,2]
##                         0 0.6206534 0.9800055
##                         1 0.4194391 0.8437833
##
##                          employment_industry
## train_b$seasonal_vaccine                  arjwrbjb     atmlpfrs     cfqqtusy
##                         0 0.4539732494 0.0317335431 0.0396887840 0.0146865985
##                         1 0.5454818975 0.0332965600 0.0286831812 0.0093270484
```

```
##                          employment_industry
## train_b$seasonal_vaccine     dotnnunm     fcxhlnwr     haxffmxo     ldnlellj
##                        0 0.0097910657 0.0626803042 0.0018358248 0.0533263397
##                        1 0.0062180323 0.1259652994 0.0104302477 0.0366061579
##                          employment_industry
## train_b$seasonal_vaccine     mcubkhph     mfikgejo     msuufmds     nduyfdeo
##                        0 0.0142494973 0.0254392867 0.0048081126 0.0139872366
##                        1 0.0063183231 0.0221642764 0.0045130880 0.0068197774
##                          employment_industry
## train_b$seasonal_vaccine     phxvnwax     pxcmvdjn     qnlwzans     rucpziij
##                        0 0.0045458519 0.0496546901 0.0005245214 0.0273625317
##                        1 0.0023066894 0.0304884164 0.0005014542 0.0091264668
##                          employment_industry
## train_b$seasonal_vaccine     saaquncn     vjjrobsf     wlfvacwt     wxleyezf
##                        0 0.0161727424 0.0270128508 0.0096162252 0.0683626191
##                        1 0.0072209407 0.0133386822 0.0067194865 0.0658910841
##                          employment_industry
## train_b$seasonal_vaccine     xicduogh     xqicxuve
##                        0 0.0455459393 0.0250021855
##                        1 0.0160465350 0.0125363554
```

```
nbPredict_b = predict(nb_b, type="class", newdata=testloc_b)
```

5.2.3 Prediction

```
# Evaluasi AUC secara lokal
(nbAUC = roc_auc(nbPredict_a, nbPredict_b))
```

```
## [1] 0.9990222
```

```
predict_a = predict(nb_a, type="raw", newdata=test)[,2]
```

```
## Warning in predict.naiveBayes(nb_a, type = "raw", newdata = test): Type mismatch
## between training and new data for variable 'h1n1_vaccine'. Did you use factors
## with numeric labels for training, and numeric values for new data?
```

```
predict_b = predict(nb_b, type="raw", newdata=test)[,2]
```

```
## Warning in predict.naiveBayes(nb_b, type = "raw", newdata = test): Type mismatch
## between training and new data for variable 'seasonal_vaccine'. Did you use
## factors with numeric labels for training, and numeric values for new data?
```

```
submission <- cbind("respondent_id"=as.numeric(rownames(test)),
                    "h1n1_vaccine"=predict_a,
                    "seasonal_vaccine"=predict_b)
head(submission)
```

```
##      respondent_id h1n1_vaccine seasonal_vaccine
## [1,]         26707 1.935398e-02     0.0458659307
```

```
## [2,]         26708 7.813051e-05     0.0001061774
## [3,]         26709 9.352826e-01     0.9842254821
## [4,]         26710 4.589260e-01     0.9872283078
## [5,]         26711 9.987114e-01     0.9883604070
## [6,]         26712 9.939014e-01     0.9982590794
```

```
# write.csv(submission,"submission_nb.csv", row.names = FALSE)
```

5.3 GBM

Catatan: perlu diubah target variable ke dalam bentuk numerik untuk gbm dengan distribusi Bernoulli di R.

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 4.1.3
```

```
## Loaded gbm 2.1.8
```

Akan dilakukan 10-fold cross validation untuk mendapatkan parameter optimal `n.trees` dan `interaction.depth`.

```
set.seed(1)
n.folds <- 10
folds_a <- createFolds(y=train_a$h1n1_vaccine, k=n.folds, list=T, returnTrain=F)
folds_b <- createFolds(y=train_b$seasonal_vaccine, k=n.folds, list=T, returnTrain=F)
```

```
try_ntrees = c(100, 150, 200, 250)
try_depths = c(11, 13, 15, 17)
```

5.3.1 Cross Validation

```
cv_AUC_a <- NULL
cv_AUC_a <- matrix(nrow = length(try_ntrees), ncol = length(try_depths))
rownames(cv_AUC_a) = try_ntrees
colnames(cv_AUC_a) = try_depths

cv_AUC_b <- NULL
cv_AUC_b <- matrix(nrow = length(try_ntrees), ncol = length(try_depths))
rownames(cv_AUC_b) = try_ntrees
colnames(cv_AUC_b) = try_depths

# Hasil cross validation tertera pada chunk berikutnya
# dikarenakan memerlukan waktu yang lama untuk mengeksekusi kode.

# tic("GBM_A CV")
# for (n in try_ntrees){
#   AUC.ave <- NULL;
#     for (d in try_depths){
#     AUC <- NULL; i=1
```

```
#     for(fold in folds_a){
#       print(paste(n,d,i)); i=i+1
#       ## GBM
#       set.seed(1)
#       mod = gbm(h1n1_vaccine~., data=train_a[-fold, ],
#                 n.trees=n, interaction.depth=d,
#                 n.minobsinnode=10, shrinkage=0.1,
#                 distribution="bernoulli", verbose=F)
#
#       ## Predicting in the validation set
#       pred = predict(mod, newdata=train_a[fold, ], type="response")
#       # print(sum(is.na(pred)))
#       pred = factor(ifelse(pred>0.5, 1, 0))
#
#
#       ## AUC
#       obs = train_a[fold, ]$h1n1_vaccine
#       ROCPred <- prediction(as.numeric(pred), as.numeric(obs))
#       auc <- performance(ROCPred, measure = "auc")
#       auc <- auc@y.values[[1]]
#
#       AUC = c(AUC, auc)
#
#       ## Freeing Memory
#       rm(mod); gc()
#     }
#     print(mean(AUC))
#     AUC.ave = c(AUC.ave, mean(AUC));
#     }
#   cv_AUC_a[paste(n), ] = AUC.ave
#   print(cv_AUC_a[paste(n), ])
# }
# toc()
#
# tic("GBM_B CV")
# for (n in try_ntrees){
#   AUC.ave <- NULL;
#     for (d in try_depths){
#     AUC <- NULL; i=1
#     for(fold in folds_b){
#       print(paste(n,d,i)); i=i+1
#       ## GBM
#       set.seed(1)
#       mod = gbm(seasonal_vaccine~., data=train_b[-fold, ],
#                 n.trees=n, interaction.depth=d,
#                 n.minobsinnode=10, shrinkage=0.1,
#                 distribution="bernoulli", verbose=F)
#
#       ## Predicting in the validation set
#       pred = predict(mod, newdata=train_b[fold, ], type="response")
#       # print(sum(is.na(pred)))
#       pred = factor(ifelse(pred>0.5, 1, 0))
#
```

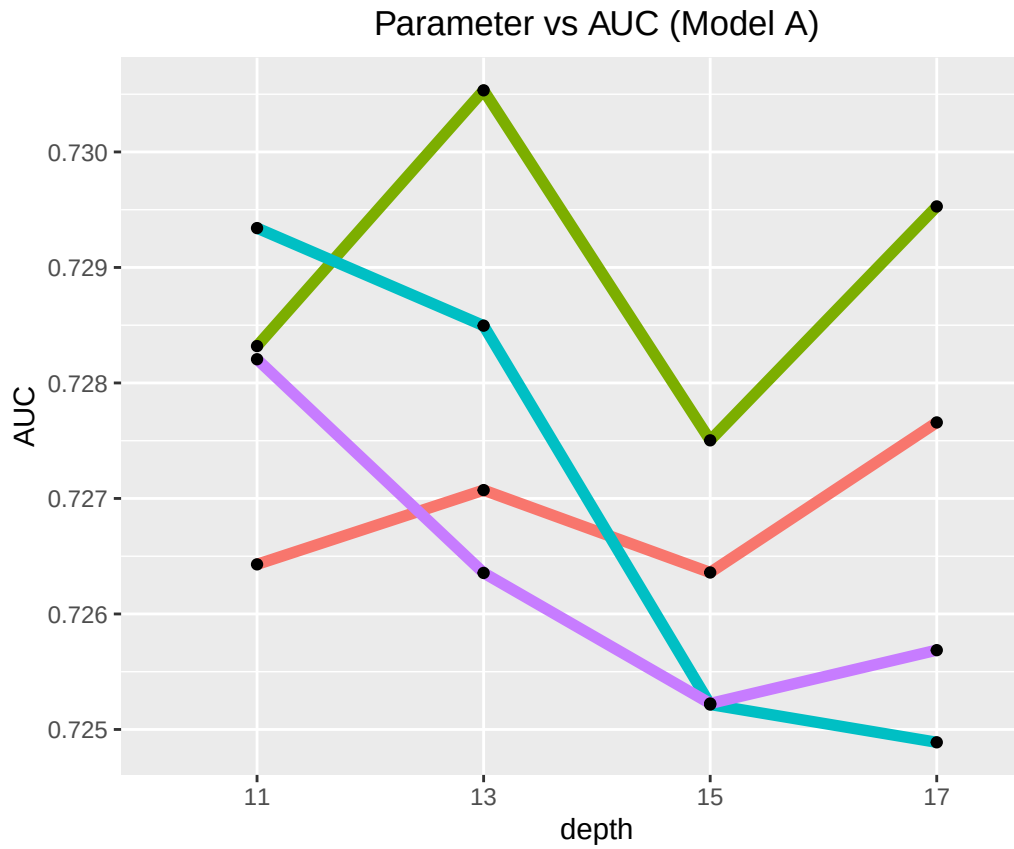
```
#
#       ## AUC
#       obs = train_b[fold, ]$seasonal_vaccine
#       ROCPred <- prediction(as.numeric(pred), as.numeric(obs))
#       auc <- performance(ROCPred, measure = "auc")
#       auc <- auc@y.values[[1]]
#
#       AUC = c(AUC, auc)
#
#       ## Freeing Memory
#       rm(mod); gc()
#     }
#     print(mean(AUC))
#     AUC.ave = c(AUC.ave, mean(AUC));
#     }
#   cv_AUC_b[paste(n), ] = AUC.ave
#   print(cv_AUC_b[paste(n), ])
# }
# toc()
```

```
library(reshape2) # melt()
```

```
## Warning: package 'reshape2' was built under R version 4.1.2
```

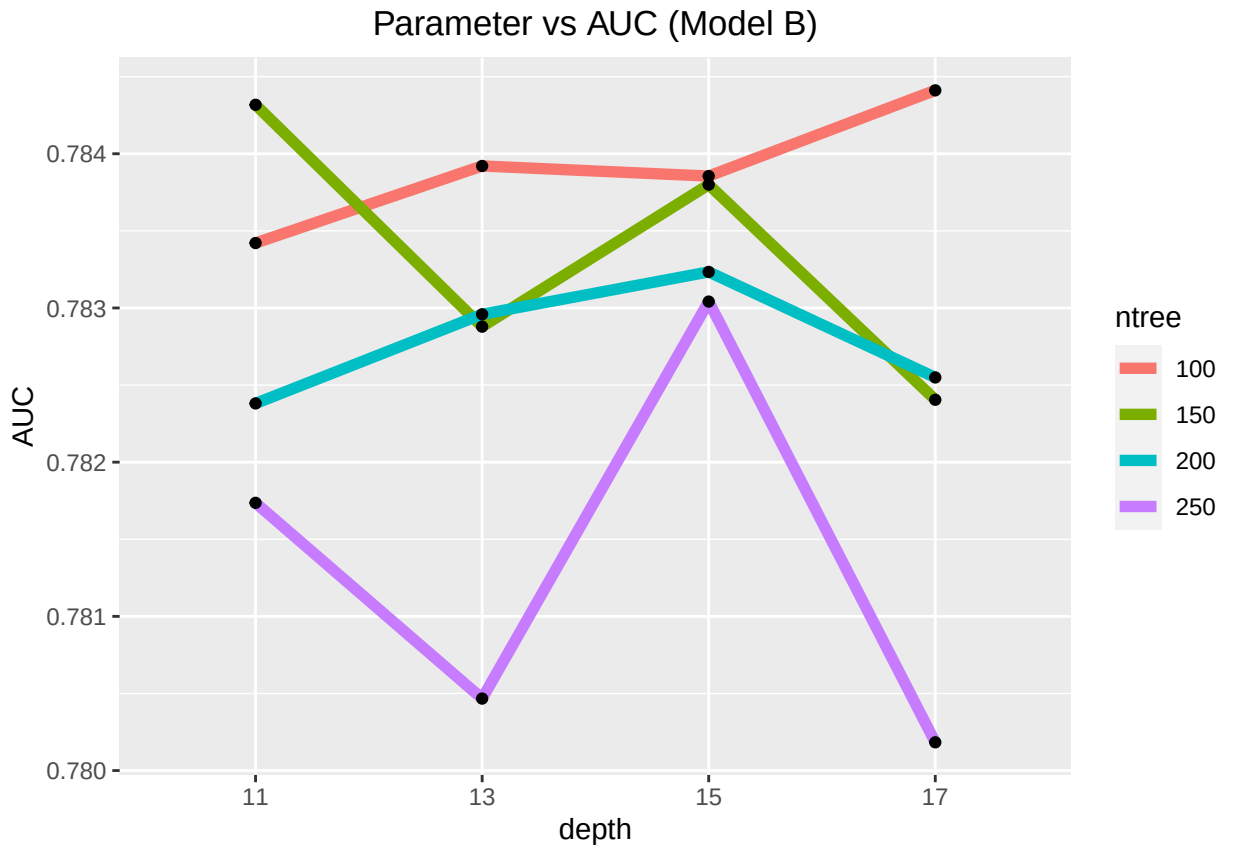
```
# Hasil Cross Validation Model A ~ time elapsed: 116.8463 mins
cv_AUC_a["100", ] <- c(0.7264294, 0.7270722, 0.7263590, 0.7276581)
cv_AUC_a["150", ] <- c(0.7283192, 0.7305335, 0.7275032, 0.7295283)
cv_AUC_a["200", ] <- c(0.7293409, 0.7284960, 0.7252160, 0.7248888)
cv_AUC_a["250", ] <- c(0.7282054, 0.7263554, 0.7252248, 0.7256865)

cv_AUC_a = melt(cv_AUC_a)
cv_AUC_a$Var1 = as.factor(cv_AUC_a$Var1)
cv_AUC_a$Var2 = as.factor(cv_AUC_a$Var2)


# Hasil Cross Validation Model B ~ time elapsed: 127.204 mins
cv_AUC_b["100", ] <- c(0.7834211, 0.7839208, 0.7838552, 0.7844113)
cv_AUC_b["150", ] <- c(0.7843176, 0.7828790, 0.7837996, 0.7824047)
cv_AUC_b["200", ] <- c(0.7823809, 0.7829589, 0.7832337, 0.7825493)
cv_AUC_b["250", ] <- c(0.7817363, 0.7804669, 0.7830414, 0.7801833)

cv_AUC_b = melt(cv_AUC_b)
cv_AUC_b$Var1 = as.factor(cv_AUC_b$Var1)
cv_AUC_b$Var2 = as.factor(cv_AUC_b$Var2)
```

```
ggplot(cv_AUC_a, aes(x = Var2, y = value)) +
  geom_line(aes(color = Var1, group = Var1), size=2) +
  geom_point()+
  ggtitle("Parameter vs AUC (Model A)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "depth", y = "AUC", color = "ntree")
```



5.3.1.1 Pemilihan Parameter

```
ggplot(cv_AUC_b, aes(x = Var2, y = value)) +
  geom_line(aes(color = Var1, group = Var1), size=2) +
  geom_point()+
  ggtitle("Parameter vs AUC (Model B)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "depth", y = "AUC", color = "ntree")
```



Tujuan pemodelan merupakan membuat model yang tidak rumit (simple). Maka, akan digunakan parameter berikut.

```
ntree_a = 150
depth_a = 13
ntree_b = 100
depth_b = 15
```

5.3.2 Model A

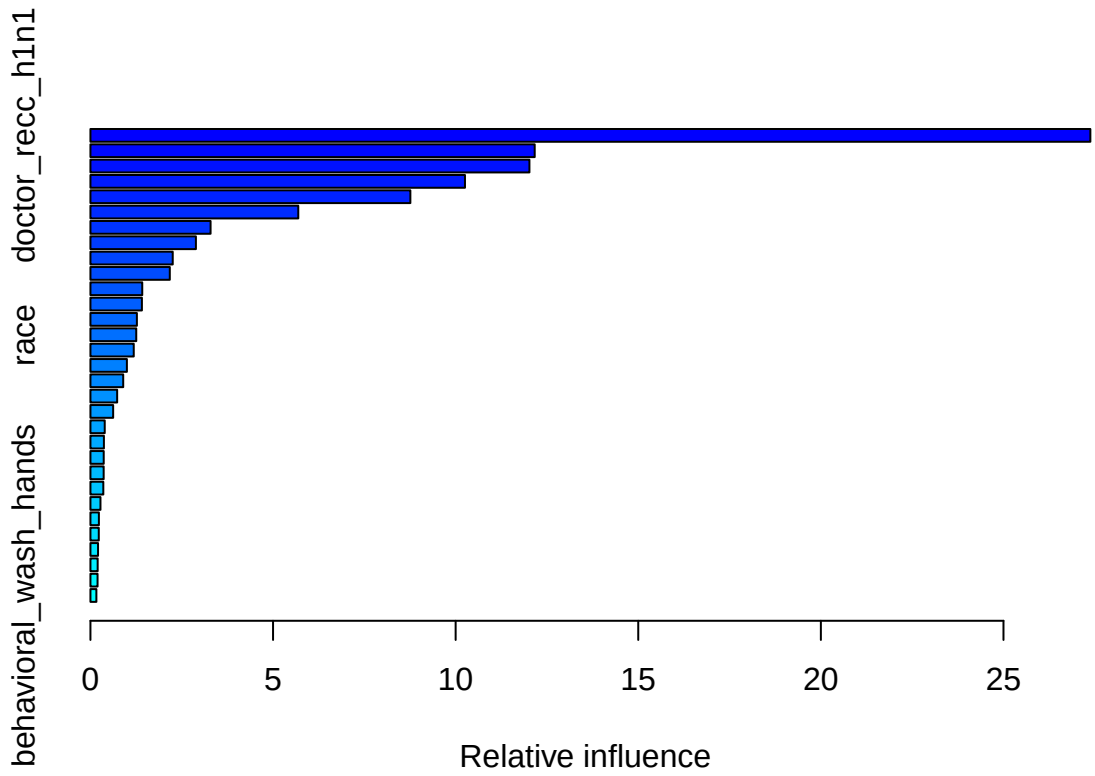
```
set.seed(1)
gbm_a = gbm(h1n1_vaccine~., data=train_a,
            n.trees=ntree_a, interaction.depth=depth_a,
            n.minobsinnode=10, shrinkage=0.1,
            distribution="bernoulli", verbose=F)
```

```
gbmPredict_a = predict(gbm_a, type="response", newdata=testloc_a)
```

```
## Using 150 trees...
```

```
gbmPredict_a = ifelse(gbmPredict_a > 0.5, 1, 0)
```

```
(imp_gbm_a <- summary(gbm_a))
```




```
##                                                        var    rel.inf
## doctor_recc_h1n1                         doctor_recc_h1n1 27.3790864
## employment_industry                   employment_industry 12.1642736
## health_insurance                         health_insurance 12.0196792
## opinion_h1n1_risk                       opinion_h1n1_risk 10.2552289
## opinion_h1n1_vacc_effective opinion_h1n1_vacc_effective  8.7600329
## hhs_geo_region                             hhs_geo_region  5.6909633
## opinion_seas_risk                       opinion_seas_risk  3.2897551
## age_group                                     age_group  2.8880863
## health_worker                              health_worker  2.2536012
## education                                     education  2.1729571
## opinion_seas_vacc_effective opinion_seas_vacc_effective  1.4185550
## h1n1_concern                                 h1n1_concern  1.4092910
## opinion_h1n1_sick_from_vacc opinion_h1n1_sick_from_vacc  1.2717437
## race                                               race  1.2540889
## income_poverty                             income_poverty  1.1860847
## h1n1_knowledge                             h1n1_knowledge  0.9983480
## census_msa                                     census_msa  0.8975211
## doctor_recc_seasonal                 doctor_recc_seasonal  0.7312707
## sex                                                 sex  0.6224055
## marital_status                             marital_status  0.3914785
## employment_status                       employment_status  0.3687240
## child_under_6_months                 child_under_6_months  0.3632624
```

```
## chronic_med_condition             chronic_med_condition  0.3622457
## behavioral_large_gatherings behavioral_large_gatherings  0.3527015
## behavioral_face_mask               behavioral_face_mask  0.2747637
## behavioral_antiviral_meds     behavioral_antiviral_meds  0.2314489
## rent_or_own                                 rent_or_own  0.2286617
## behavioral_outside_home         behavioral_outside_home  0.2079994
## behavioral_avoidance               behavioral_avoidance  0.1979146
## behavioral_touch_face             behavioral_touch_face  0.1947710
## behavioral_wash_hands             behavioral_wash_hands  0.1630560
```

5.3.3 Model B

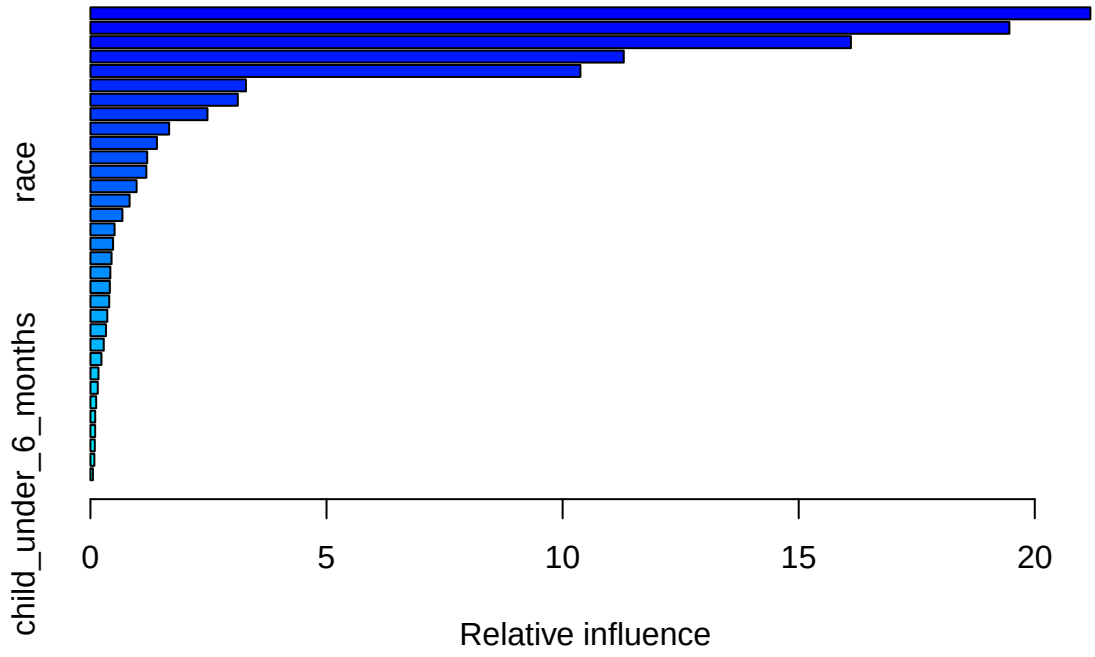
```
set.seed(1)
gbm_b = gbm(seasonal_vaccine~., data=train_b,
            n.trees=ntree_b, interaction.depth=depth_b,
            n.minobsinnode=10, shrinkage=0.1,
            distribution="bernoulli", verbose=F)
```

```
gbmPredict_b = predict(gbm_b, type="response", newdata=testloc_b)
```

```
## Using 100 trees...
```

```
gbmPredict_b = ifelse(gbmPredict_b > 0.5, 1, 0)
```

```
(imp_gbm_b <- summary(gbm_b))
```




```
##                                                  var     rel.inf
## opinion_seas_vacc_effective opinion_seas_vacc_effective 21.17677936
## opinion_seas_risk                     opinion_seas_risk 19.46159567
## doctor_recc_seasonal               doctor_recc_seasonal 16.10585678
## employment_industry                 employment_industry 11.29498473
## age_group                                     age_group 10.37688825
## hhs_geo_region                           hhs_geo_region  3.29324521
## opinion_seas_sick_from_vacc opinion_seas_sick_from_vacc  3.12109132
## health_insurance                       health_insurance  2.47820126
## education                                     education  1.66732492
## health_worker                           health_worker  1.40823900
## h1n1_knowledge                         h1n1_knowledge  1.20237360
## race                                             race  1.18436295
## opinion_h1n1_risk                   opinion_h1n1_risk  0.97762985
## income_poverty                         income_poverty  0.82979482
## rent_or_own                               rent_or_own  0.67733010
## opinion_h1n1_vacc_effective opinion_h1n1_vacc_effective  0.51110719
## h1n1_concern                             h1n1_concern  0.47974837
## employment_status                   employment_status  0.44721036
## doctor_recc_h1n1                     doctor_recc_h1n1  0.42143188
## behavioral_touch_face           behavioral_touch_face  0.41086641
## chronic_med_condition           chronic_med_condition  0.39744270
## census_msa                                 census_msa  0.35738164
## household_children                 household_children  0.32853590
## marital_status                         marital_status  0.28258805
## household_adults                     household_adults  0.23259454
```

```
## sex                                                 sex  0.17036182
## behavioral_antiviral_meds     behavioral_antiviral_meds  0.15488720
## behavioral_face_mask               behavioral_face_mask  0.12157371
## behavioral_outside_home         behavioral_outside_home  0.10099459
## behavioral_large_gatherings behavioral_large_gatherings  0.09988038
## behavioral_wash_hands             behavioral_wash_hands  0.09178009
## behavioral_avoidance               behavioral_avoidance  0.08314906
## child_under_6_months               child_under_6_months  0.05276830
```

```
# write.csv(head(imp_gbm_a),"imp_gbm_a.csv", row.names = FALSE)
# write.csv(head(imp_gbm_b),"imp_gbm_b.csv", row.names = FALSE)
```

5.3.4 Prediction

```
# Evaluasi AUC secara lokal
(gbmAUC = roc_auc(gbmPredict_a, gbmPredict_b))
```

```
## [1] 0.787086
```

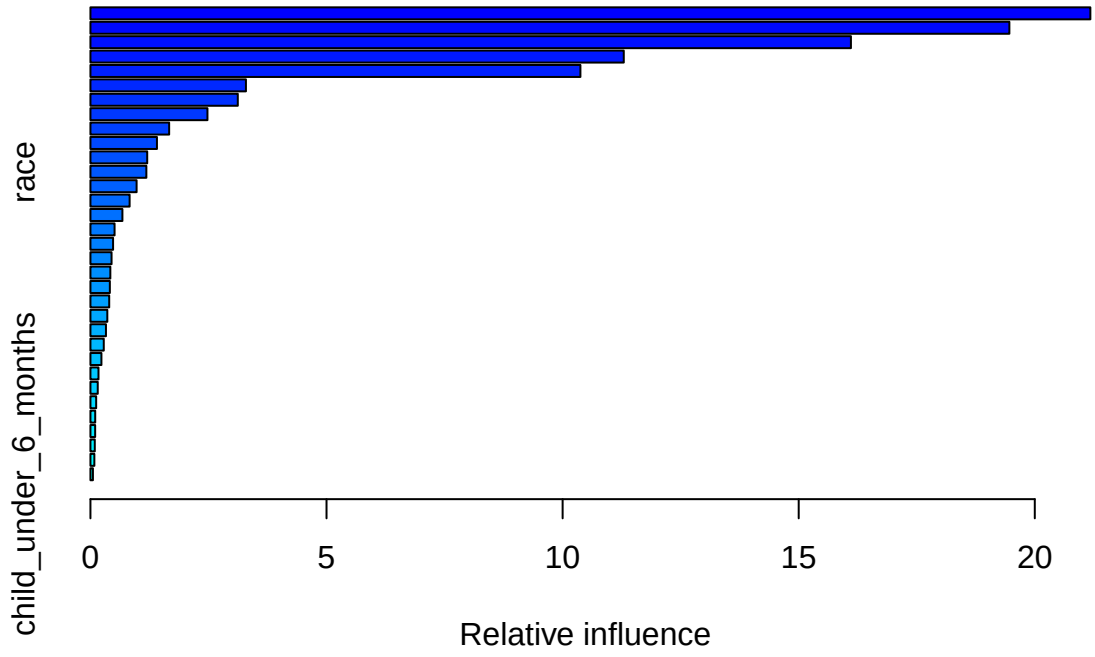
```
predict_a = predict(gbm_a, type="response", newdata=test)
```

```
## Using 150 trees...
```

```
predict_b = predict(gbm_b, type="response", newdata=test)
```

```
## Using 100 trees...
```

```
(imp_gbm_b <- summary(gbm_b))
```




```
##                                                    var     rel.inf
## opinion_seas_vacc_effective opinion_seas_vacc_effective 21.17677936
## opinion_seas_risk                     opinion_seas_risk 19.46159567
## doctor_recc_seasonal               doctor_recc_seasonal 16.10585678
## employment_industry                 employment_industry 11.29498473
## age_group                                     age_group 10.37688825
## hhs_geo_region                           hhs_geo_region  3.29324521
## opinion_seas_sick_from_vacc opinion_seas_sick_from_vacc  3.12109132
## health_insurance                       health_insurance  2.47820126
## education                                     education  1.66732492
## health_worker                           health_worker  1.40823900
## h1n1_knowledge                         h1n1_knowledge  1.20237360
## race                                             race  1.18436295
## opinion_h1n1_risk                   opinion_h1n1_risk  0.97762985
## income_poverty                         income_poverty  0.82979482
## rent_or_own                               rent_or_own  0.67733010
## opinion_h1n1_vacc_effective opinion_h1n1_vacc_effective  0.51110719
## h1n1_concern                             h1n1_concern  0.47974837
## employment_status                   employment_status  0.44721036
## doctor_recc_h1n1                     doctor_recc_h1n1  0.42143188
## behavioral_touch_face           behavioral_touch_face  0.41086641
## chronic_med_condition           chronic_med_condition  0.39744270
## census_msa                                 census_msa  0.35738164
## household_children                 household_children  0.32853590
## marital_status                         marital_status  0.28258805
## household_adults                     household_adults  0.23259454
```

```
## sex                                                 sex  0.17036182
## behavioral_antiviral_meds     behavioral_antiviral_meds  0.15488720
## behavioral_face_mask               behavioral_face_mask  0.12157371
## behavioral_outside_home         behavioral_outside_home  0.10099459
## behavioral_large_gatherings behavioral_large_gatherings  0.09988038
## behavioral_wash_hands             behavioral_wash_hands  0.09178009
## behavioral_avoidance               behavioral_avoidance  0.08314906
## child_under_6_months               child_under_6_months  0.05276830
```

```
submission <- cbind("respondent_id"=as.numeric(rownames(test)),
                    "h1n1_vaccine"=predict_a,
                    "seasonal_vaccine"=predict_b)
head(submission)
```

```
##      respondent_id h1n1_vaccine seasonal_vaccine
## [1,]         26707   0.11650236       0.32845188
## [2,]         26708   0.02785604       0.04255647
## [3,]         26709   0.07064625       0.51133433
## [4,]         26710   0.82372337       0.89757263
## [5,]         26711   0.41544670       0.54139925
## [6,]         26712   0.88400249       0.96272045
```

```
# write.csv(submission,"submission_gbm.csv", row.names = FALSE)
```

6 Hasil Akhir

```
data.frame("Model"=c("Regresi Logistik", "Naive Bayes", "GBM"),
           "AUC Validasi Lokal" = round(c(logrAUC, nbAUC, gbmAUC),4),
           "AUC Pengumpulan" = c(0.8505, 0.7973, 0.8548))
```

```
##              Model AUC.Validasi.Lokal AUC.Pengumpulan
## 1 Regresi Logistik             0.7482          0.8505
## 2      Naive Bayes             0.9990          0.7973
## 3              GBM             0.7871          0.8548
```