

Relating SARS-Cov-2 Infection Risks To Relative Risks of Dietary Items in Different Ethnicities By Machine Learning

Mohammad Reza Besharati, PhD Candidate, Sharif University of Technology, Tehran, Iran
Nafiseh Jafari, Graduate Student, Malek Ashtar University of Technology, Tehran, Iran
Dr. Mohammad Izadi, Associate Professor, Sharif University of Technology, Tehran, Iran
Dr. Alireza Talebpour, Associate Professor, Shahid Beheshti University, Tehran, Iran
Dr. Maryam Hourali, Assistant Professor, Malek Ashtar University of Technology, Tehran, Iran

If the risk of SARS-Cov-2 is related to the diet of different people of different ethnicities, then we naturally expect the following equation holds for all ethnicities such as E:

$$(DietRegime_E) \times (DietEffect_E) = COVIDStatus_E$$

By DietEffect, we mean the set of “Relative Risks” of different Dietary Items (98 Items), calculated for each ethnicity separately (27 different Ethnicities).

For Dietary Item I we have:

$$RelativeRisk(I) = \frac{P(AparentInfection | Consume I)}{P(AparentInfection | not Consume I)}$$

By COVIDStatus, we mean the “Relative Risks” of Ethnicities themselves. For Ethnicity E we have:

$$RelativeRisk(E) = \frac{P(AparentInfection | E)}{P(AparentInfection | not E)}$$

DietRegime variable is unknown. But the variables DietEffect and CovidStatus can be calculated from the collected data (more than 16000 family records, Anonymous self-reported data), for all ethnicities such as E. Then, a machine Learning algorithm could be used to relates DietEffect (= features for Learning Process) to COVIDStatus (= Class Label for classification)

In order to create a Training Dataset with Size=N, we randomly selected a subset of Data Records (with 10000 members in each subset), N times. We then calculated the DietEffect and COVIDStatus values for each subset for all Ethnicities.

By applying different machine Learning algorithms, we noticed that the above equation holds for 27 different ethnicities, with a somehow good accuracy 85% (See table-1, Calculated by 10 fold cross validation). So we could reason that the diet has a key role in Risks of apparent SARS-Cov-2.

Table-1

Algorithm	N	Accuracy %	F-Measure %	ROC Area
Multilayer Perceptron	27000	75.3	75.3	0.80
LibSVM	27000	82.85	82.85	0.82
RandomCommittee	270000	82.78	82.78	0.91
Bagging	270000	83.05	83.05	0.91
RandomSubSpace	270000	83.82	83.82	0.92
RandomForest	270	74.8	74.8	0.82
RandomForest	2700	81.1	81.1	0.89
RandomForest	27000	82.81	82.8	0.91
RandomForest	270000	84.91	84.91	0.93
RandomForest	493772	85.30	85.30	0.93

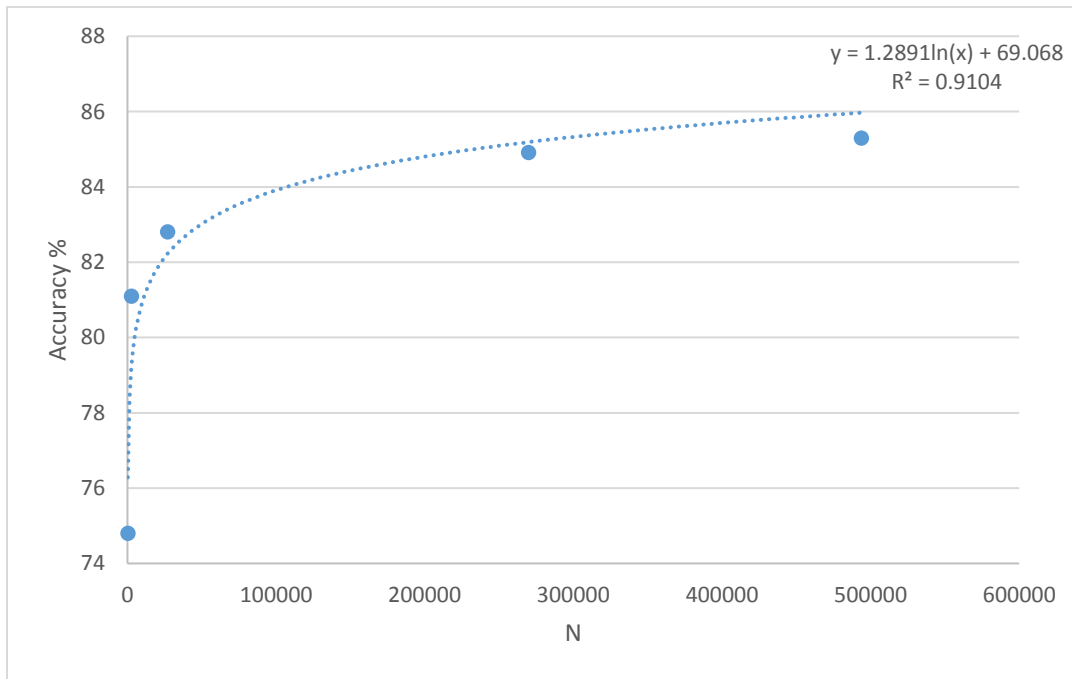


Figure 1- Accuracy for RandomForest Classification Algorithm, Calculated by 10 fold cross validation.