Confidentiality Notice :
Non-confidential report

CoronaSurveys
Estimating Active Cases of COVID-19

Author :
KACEM Mohamed

Promotion :
2023

2re année Mathématiques appliquées

ENSTA Paris Tutor :
Andrea SIMONETTO

Host Organism Tutor :
Antonio FERNÁNDEZ ANTA

Internship from 12 May 2022 to 12 August 2022

Name of the host organism : IMDEA Networks Institute

Adress : Avenida del Mar Mediterráneo 22, Leganés, Comunidad de Madrid 28918

Acknowledgments

I would like to thank some people who made this internship possible. I sincerely acknowledge and appreciate their contribution and assistance.

Many thanks to Professor Antonio Fernández Anta, my supervisor at IMDEA, for guiding me during this entire project. His advices helping me choosing the internship subject and leading me to relevant and interesting topics are deeply appreciated.

To Jesús Rufino and Juan Ramírez who were always there for me to help me answer all my questions as well as all the other colleagues whose articles were with a huge help to me.

To Professor Andrea Simonetto, researcher and teacher at the Applied Mathematics Department of ENSTA Paris, who accepted to be my referent teacher and who was always available for me.

Abstract

Covid with all its variants has changed our daily life considerably, which makes it necessary to study this pandemic to better understand its mechanisms and influence, and to have effective methods to diagnose or predict it even in places or/and times where screening tools are not available. In this project, we will use data collected in the Global COVID-19 Trends and Impact Surveys (CMU Global CTIS) to estimate the number of people affected by covid in different states of the United States among people with different vaccination status using different models and methods, first based directly on participants' responses through symptoms and finally applying machine learning classification models. The probability of catching the covid being unvaccinated is almost 10 times more, but the effect of the vaccine is different from one period to another (decrease with time which justifies the interest of doing the booster doses). The different classification models give satisfactory results for the estimates made, the best performances in terms of correlation and precision are Random Forest and XGBoost, this is confirmed through the Performance Metrics such as the F1 Score as well as through the ROC curves.

Key words

— Classification
— Regression
— Infection
— Prevalence
— Random Forest
— XGBoost
— SVM
— covid-19
— United-States

Table des figures

Table des matières

1 The framework of the project

1.1 Introduction

The world is suffering from a pandemic called COVID-19, caused by the SARS-CoV-2 virus that, since its first appearance at the end of 2019, has had a huge increase all over the world. And over time it has developed more dangerous variants with stronger immunity. When testing availability and other resources are limited, national and local governments can encounter problems evaluating the reach and evolution of the epidemic. Hence, any means to evaluate the evolution of the pandemic and its impact with reasonable level of accuracy is useful [1].
And even though vaccination has shown considerable effectiveness in protecting against the pandemic, the results of several studies show that this effectiveness is decreasing, especially after 3 months from the last dose [2].
Since the year 2020, CMU (Carnegie Mellon University) has collaborated with Facebook to collect large research databases on behaviors, symptoms, infection, testing and, more recently, vaccination status of participants (CMU Global CTIS). We will use some of these combinations and test their accuracy and precision by applying them to data collected from different states in the U.S.
In this study we will also use the portion of self-reported confirmed cases infected with covid-19 from a subset of the CMU Global CTIS survey responses, in order to develop improved models of classification. Several classifiers will be used (Logistic regression, Random Forest, XGBoost, linear SVM and Neural Network)

1.2 Receiving the data

The data collection starts from a survey of some direct and indirect questions, called the COVID-19 Trends and Impact Surveys (CTIS) led by Carnegie Mellon University (CMU) in a collaboration with Facebook. IMDEA Networks has signed Data Use Agreements with Facebook and Carnegie Mellon University (CMU) to access their data.
The next step is the processing of the collected databases in order to use it for this project. Our goal is to produce a pipeline that will make us able to extract the results we want about the infections of Covid19 and the vaccine efficiency using different methods to be able the compare these methods.

1.3 Steps of data processing

Before starting our work on the databases, we notice that the number of questions is huge, so we have so many columns, and some questions are not important. this huge number of features will make the learning and the training take forever and can even negatively affect the results, so we have to keep just a certain number of features.
In order to reduce the number of columns, we have chosen some of the direct questions that affect the results the most and which is related to symptoms, infections and vaccination status.
Thus, we choose to keep the following columns : EndDatetime, fips, A1_1, A1_2, A1_3, A1_4, A1_5, A5_1, A5_2, A5_3, A2, B2b, B10c, V1, V2, D1, D2, C14, C14a, C14b, C14c, B13a, B10 , B2, B2_14_TEXT which correspond to :

- EndDatetime : indicates the date
- A1 : you or anyone in your household experienced symptoms.
- B2 : have you personally experienced any of the following symptoms ?
- A5 : How many people, including you, are currently staying in your household (1, 2, 3 according to age)
- B2b : how many days have you had at least one new or unusual symptom.
- B10 : you been tested for COVID-19 in the past 14 days.
- B10c : your most recent test find that you have COVID-19.
- V : vaccination.
- D : gender and age.
- C14 : C14a In the past 7 days, how often did you wear a mask when in public ?

— B13a : Have you ever had coronavirus (COVID-19) ?

Then, we move on to the data processing which goes through the following steps :

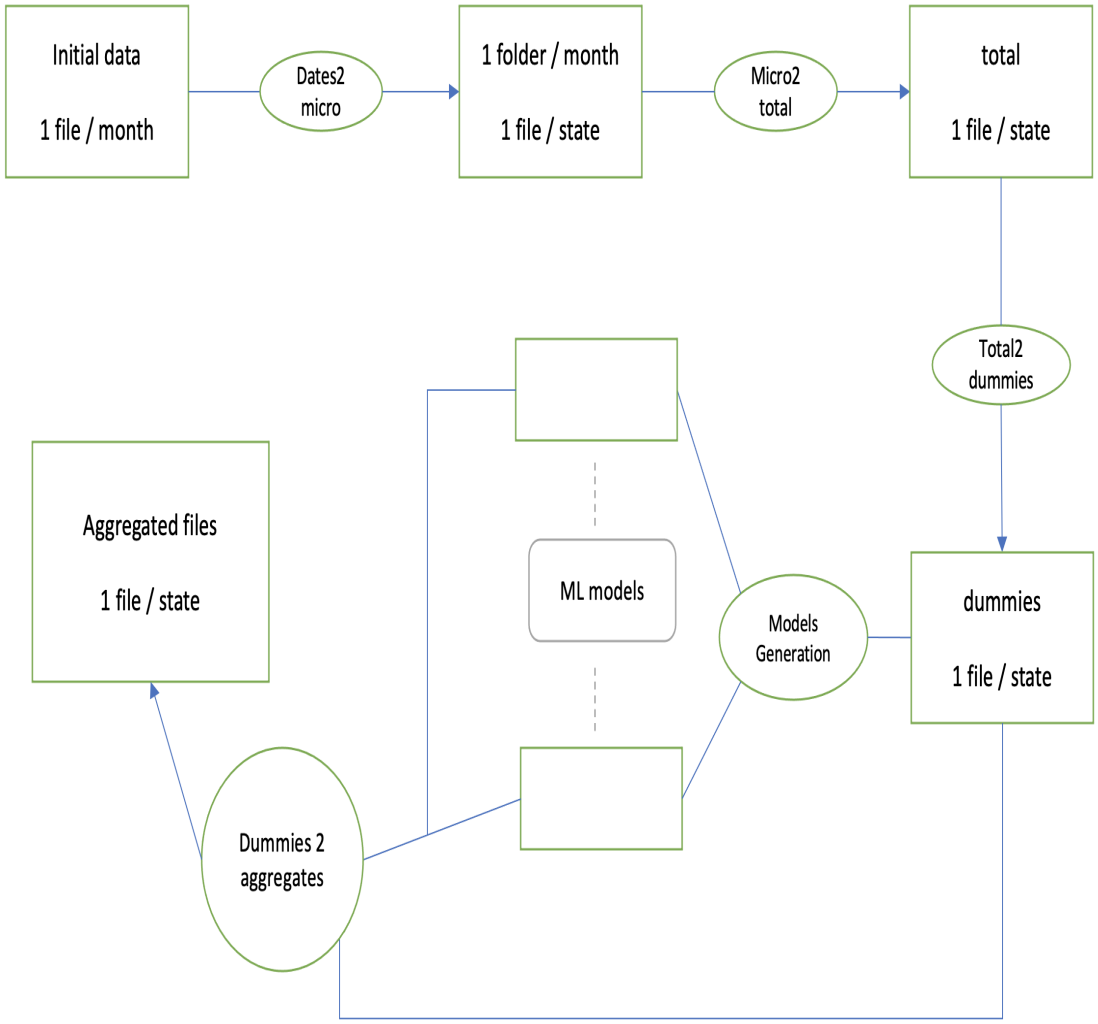

FIGURE 1.1 – Data processing

— Dates2micro : In this step we take the initial data which are in the form of comma-separated values (CSV) files representing the participants' responses, we have one file per month. This step consists of keeping only the characteristics we mentioned before and creating one file per month, each file contains

one file per state.

— Micro2total : Now we take the microdata provided by the last step and merge the files from all months for each state, thus creating the total files (one file per state). Each file therefore contains all the responses corresponding to that state from the beginning of the survey to its end.

— Total2dummies : At this point, we eliminate responses that seem irrational and illogical. Participants who do not answer seriously or who intentionally want to ruin the survey. For example, we set limits to eliminate responses with very large or negative values, or people who report having all the symptoms (there are 20 different symptoms)
Then we proceed to the dummification which consists in giving a table with binary values (0 or 1). For example for a question Q that can have three different answers 1, 2 or 3, we create 3 columns Q.1, Q.2 and Q.3. If for example Q = 2, then Q.1=0, Q.2=1 and Q.3=0.

— Dummies2aggregates : This is one of the last steps of the project and contains the results that we will try to highlight.
It consists in creating the files giving the different estimates using the different approaches either through direct combinations based on the answers, or by creating learning models based on the ground-truth.
But these results are given in a grouped way in order to be able to share the results, according to the data use contract between IMDEA, CMU and Facebook.
The results will be grouped by at least 100 answers per line. We will try to group them by day if we have more than 100 answers and we add a day in case of exception (we can find exceptions for the states with the lowest population).

2 Methods of classification

2.1 Self-reported Survey Data

2.1.1 The ground-truth

After receiving and processing the data, the next step is to determine which responses correspond to positive COVID19 cases. This task can be done directly by considering participants who answered positively to questions B10 : "Have you been tested for COVID-19 in the last 14 days ?" and who have a positive or negative answer to question B10c : "Did your most recent test show that you have COVID-19 ?". Thus, a participant who answered positively to both questions is considered an infected case and a participant who answered positively to question B10 and negatively to question B10c is considered an uninfected case. This is a set of already classified responses that we call the ground-truth, and it gives a direct and available answer to our question about infection status. Thus, to estimate the number of infections in each state, we could simply calculate the fraction of infected people among those who took a covid test. This is the so-called *test positivity rate* (TPR) and it is known to be usually higher than the actual infection rate.

2.1.2 CLI : COVID-like illness classifiers

We will not be able to use the ground-truth directly to obtain results on the populations of U.S. This approach would not be very reliable because this set is usually very small and fixed (it is not generated by random sampling), and the people who get tested are mostly people who think they have reason to be infected. Thus, calculating the fraction of infected people in a state just by going through the ground-truth will not give good results.
In this work we use some methods that also use the participants' responses directly but through more developed combinations based on symptoms called CLI [3] criterion (COVID-like illness classifiers).
Here we use two, CMU CLI which uses participants' self-reported symptoms, and Local CLI which uses indirect responses from participants who report on the conditions and symptoms of people around them.

— CMU CLI : A response is considered to be assumed to be positive if it affirms having fever (symptom B2.1) + an other symptom.

— CLI-local (indirect reported data) : CLI in household based in the variable A2 giving symptoms for people in household.

To make our predictions and generalize them to all states, we calculate the fractions in order to have normalized values. We also determine the confidence intervals that we add on the curves in order to have an idea about the accuracy and precision of the values found.

We take the example of California, which is the most populous state, and plot the fraction curves for both methods with their confidence intervals.

— For CMU CLI :

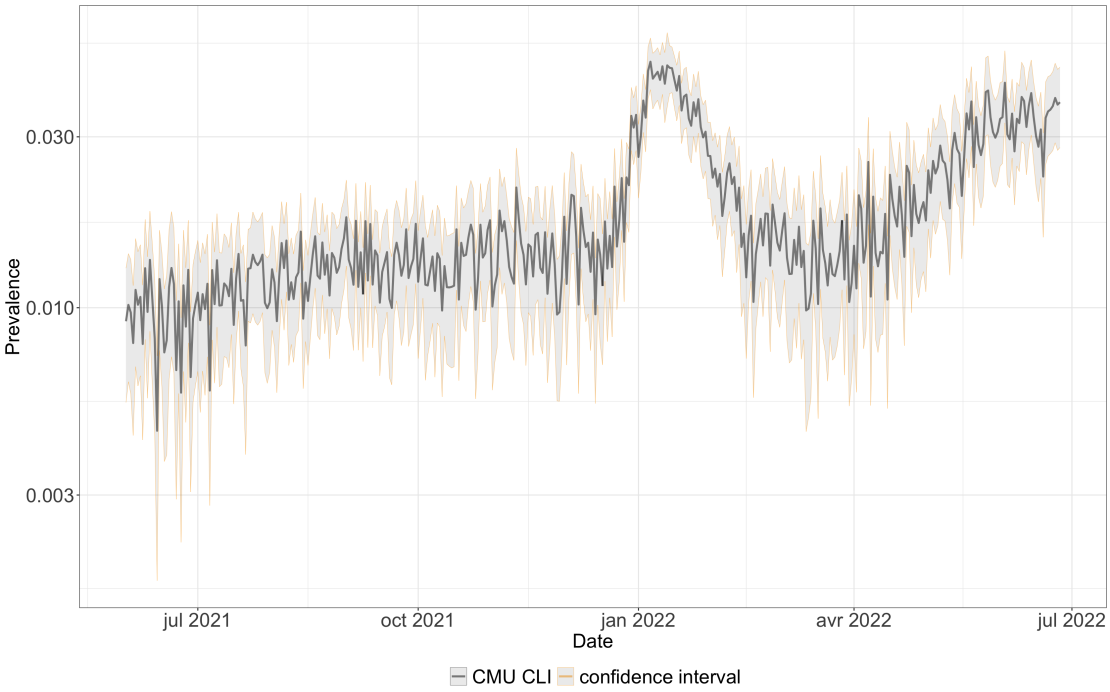

FIGURE 2.1 – Prevalence in California using CMU CLI classifier

— For CLI Local :

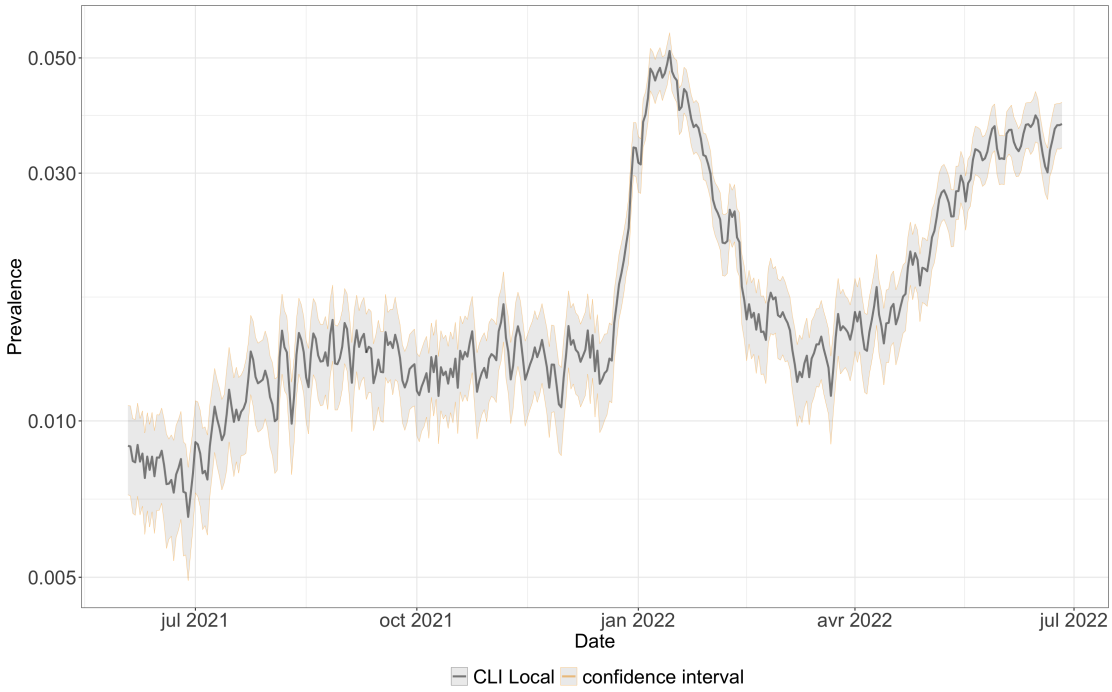

FIGURE 2.2 – Prevalence in California using CLI Local classifier

2.2 Machine learning classifiers

In our survey we have many question that we already selected certain of them. For our classification we use only questions that have discrete values, so we remove the two questions that helped us to create the ground-truth.
We use different classification models having algorithms with various constructs and compliances : Logistic regression, XGBoost, random forest, linear SVM and Neural Network.

2.2.1 Logistic regression

Logistic regression is a statistical model for studying the relationships between a set of categorical variables X and a qualitative variable Y. It is a generalized linear model using a logistic function as a link function.
A logistic regression model can also predict the probability of an event occurring (value of 1) or not (value of 0) based on the optimization of the regression coefficients. This result always varies between 0 and 1. When the predicted value is above a threshold, the event is likely to occur, while when this value is below the same threshold, it does not occur [12].

For our logistic regression we choose a binomial distribution with a complete model taking into account all the parameters we choose to keep.
We use the California example as before. Let us plot the evolution of infections for this state by logistic regression adding the confidence interval :

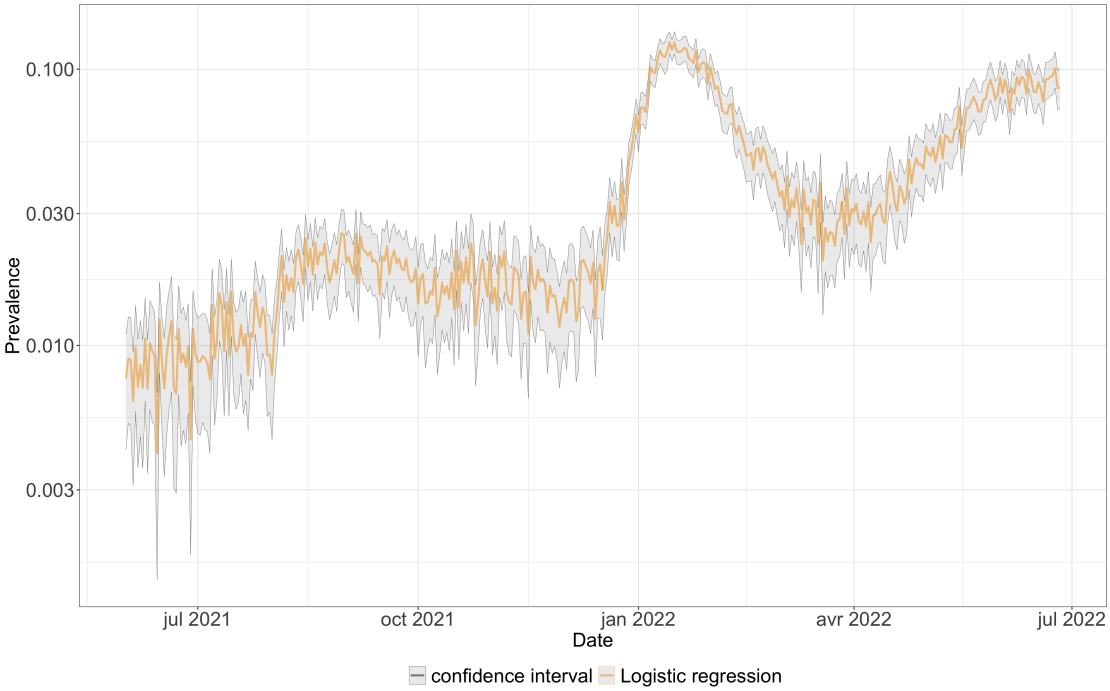

FIGURE 2.3 – Prevalence in California using logistic regression with confidence interval

2.2.2 XGBoost

XGBoost is in fact a particular version of the Gradient Boost algorithm. Indeed, it is an assembly of "weak learners" who predict the residues, and correct the errors of the previous "weak learners".
The particularity of XGBoost lies in the type of "weak learner" used. The "weak learners" are decision trees. Trees that are not good enough are "pruned", i.e. branches are cut off, until they are good enough. Otherwise they are completely removed. This method is called "pruning".
This way, XGBoost ensures that only good weak learners are kept.
In addition, XGBoost is computer optimized to make the various calculations required to apply a Gradient Boosting fast.
Finally, XGBoost offers a very large panel of hyperparameters. Thanks to this

diversity of parameters, it is possible to have total control over the implementation of Gradient Boosting.
We apply XGBoost on our examples using the ground-truth for training and we generate the estimate of the number of positive cases per day with for all states, as well as its fraction compared to the population and we generate also the confidence interval [13].
We take the California example and plot the prevalence with its confidence interval :

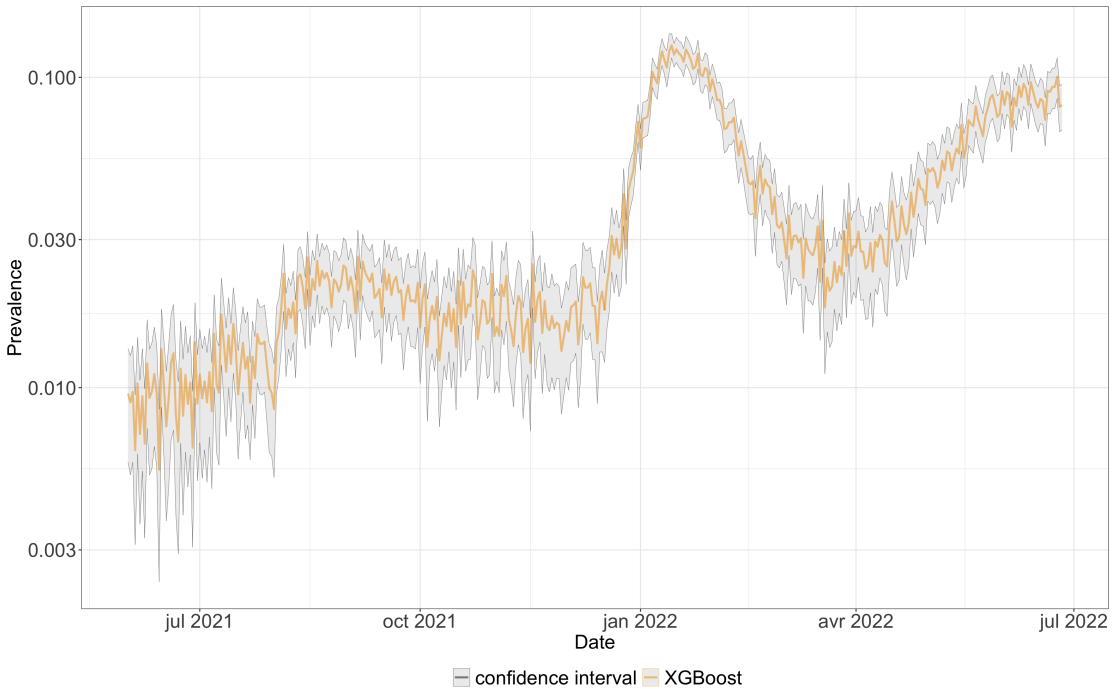

FIGURE 2.4 – Prevalence in California using XGBoost with confidence interval

2.2.3 Random Forest

Random Forest is an easy-to-interpret, stable technique that generally has good accuracies and can be used for regression or classification tasks.
With the word Forest we understand that this model is built with several trees called decision trees.
Decision trees, as their name indicates, help to make a decision based on a series of questions having a binary answer (yes or no) to lead to the final answer.

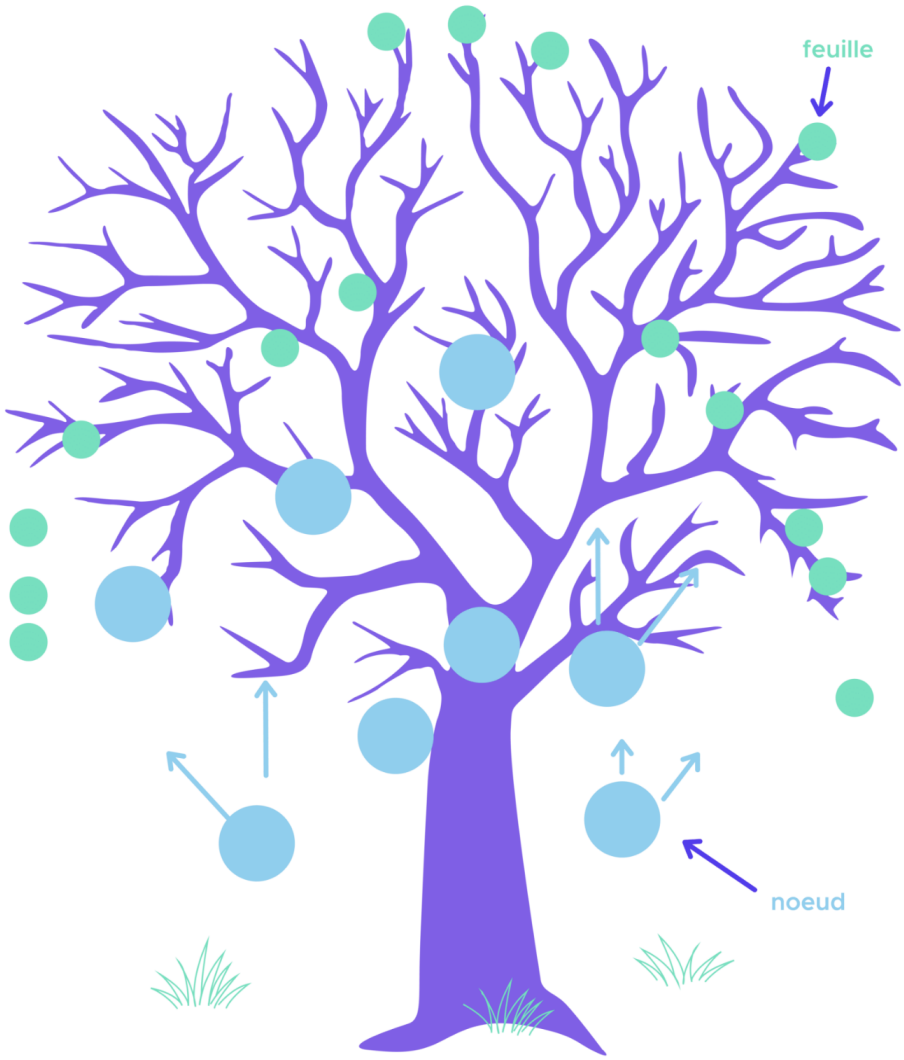

FIGURE 2.5 – Decesion tree [11]

On the tree each question corresponds to a node separating two branches, depending on the answer we go to one of its two branches and we continue like this until we arrive at one of the leaves of the tree and which contains the answer.
The Random Forest contains several decision trees that each contain an answer, each decision tree has a contribution similar to a vote either through the average of the results (regression) or by looking at the most frequent answer in the trees (Classification) to give the final answer [11].

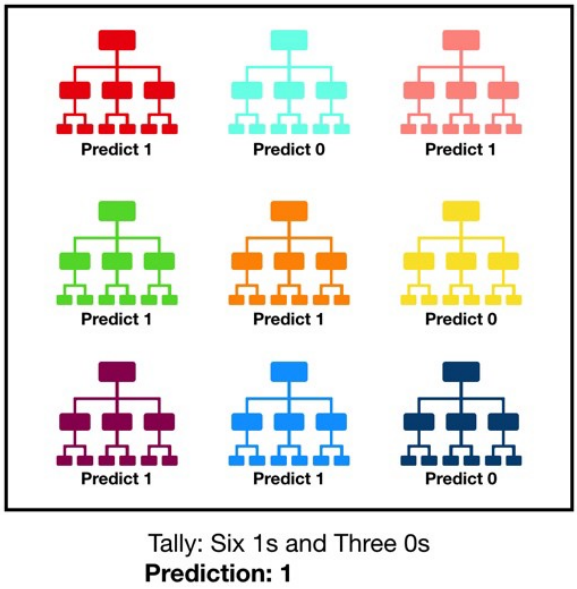

FIGURE 2.6 – Random forest [11]

We apply now two Random forest for different states to get the fraction of infected people and we calculate the confidence intervals.
We take the example of California as before. Let's plot the evolution of the prevalence for this state by Random Forest adding the confidence interval :

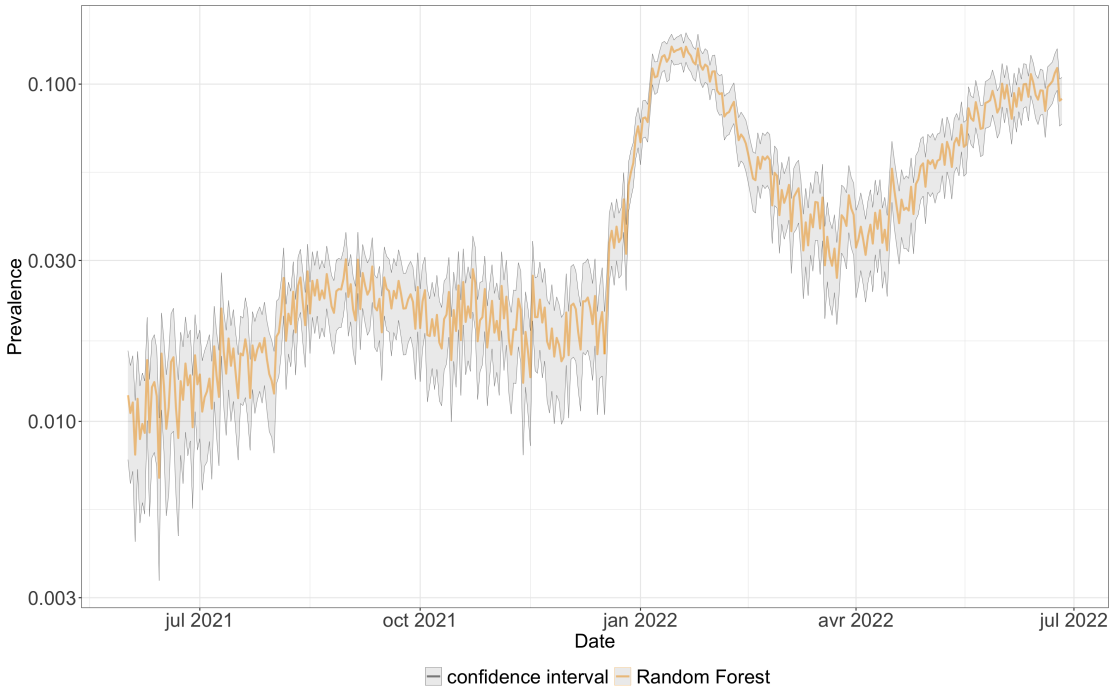

FIGURE 2.7 – Prevalence in California using Random Forest with confidence interval

2.2.4 Support Vector Machine

SVMs are classifiers that allow to treat nonlinear problems by reformulating them into quadratic optimisation problems. Which are much easier to solve. Concretely, it consists in finding, in a space of dimension $N > 1$, the hyperplane that best divides a dataset into two. For linear SVM, the boundary separating the classes is a straight line.
We do our predictions using linear SVM and we plot the evolution of the number of infections per day and we add the confidence interval :

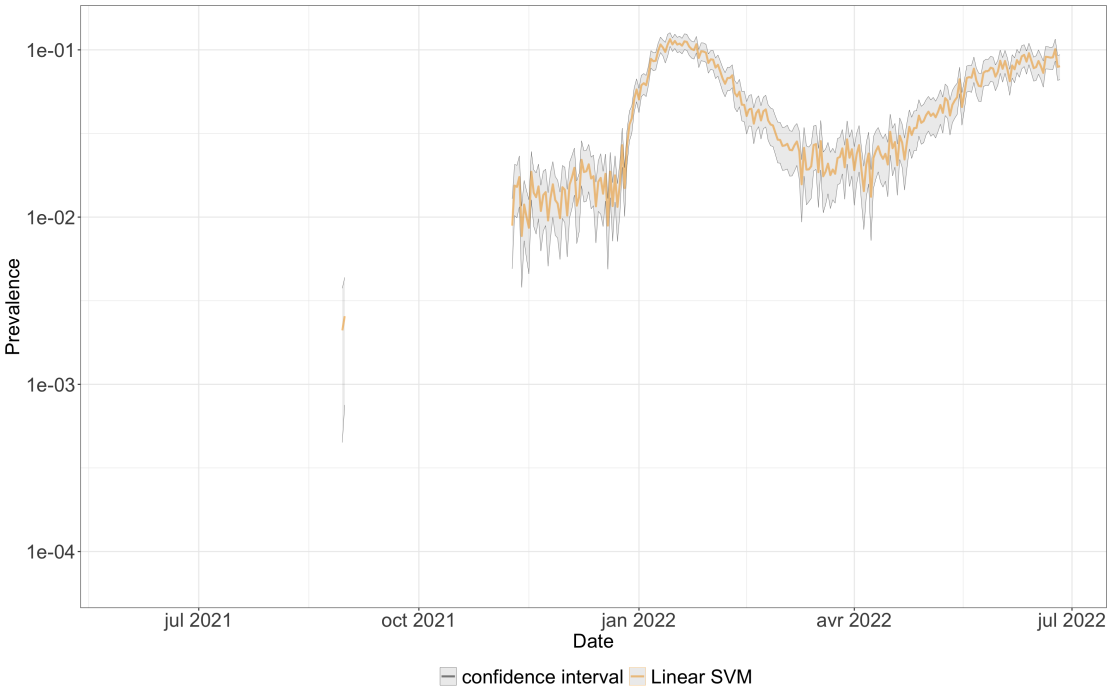

FIGURE 2.8 – Prevalence in California using Linear SVM with confidence interval

The first part missing of the curve is missing due to NA values, and because of some values that we had to change to avoid some errors, but it give good approximation to the infections number evolution.

2.2.5 Neural Network

A neural network consists of :
- Input layers : Layers that take inputs based on existing data.
- Hidden layers : Layers that use back propagation to optimise the weights of the input variables in order to improve the predictive power of the model.
- Output layers : Output of predictions based on the data from the input and hidden layers [10].

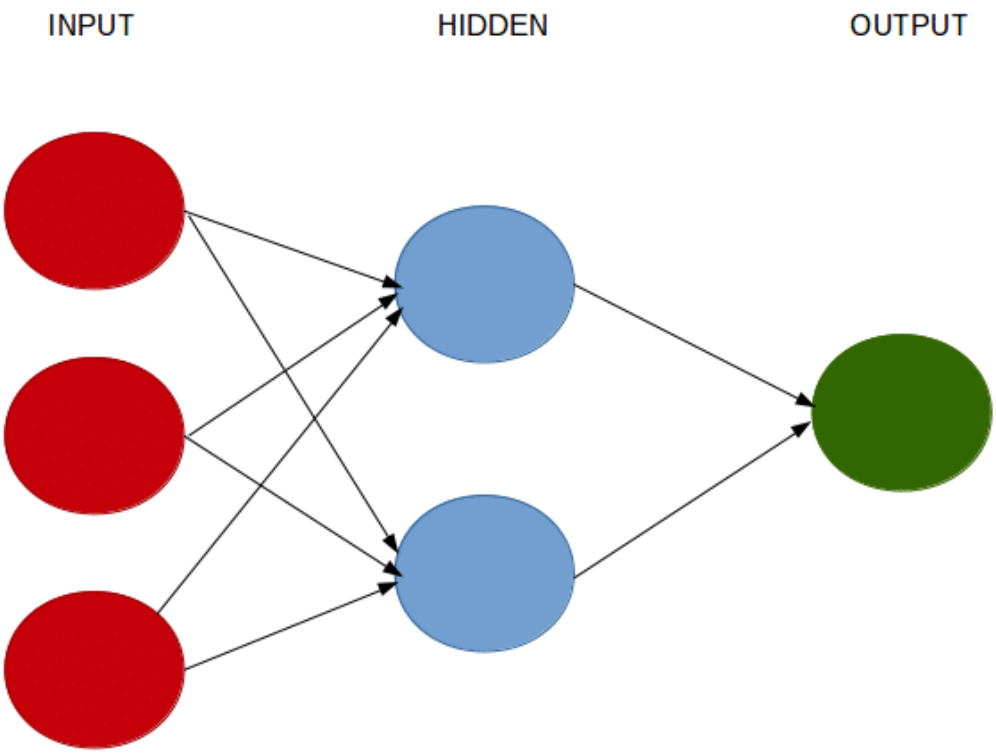

FIGURE 2.9 – Neural networks [10]

For Neural network we do the same plot as before for California and we get :

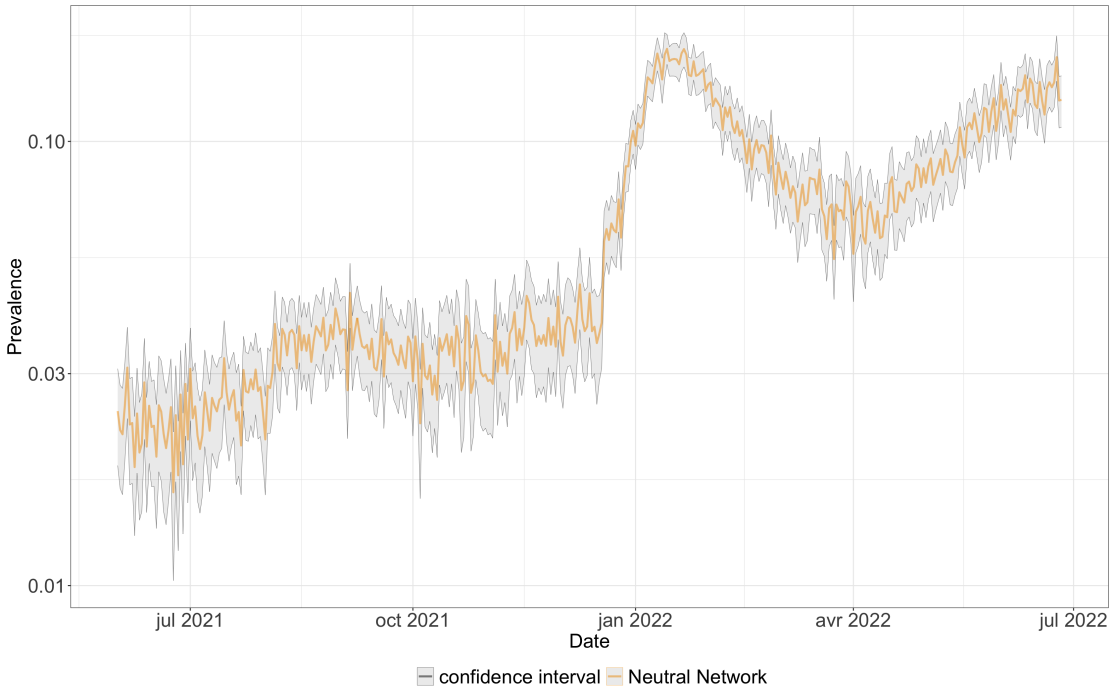

FIGURE 2.10 – Prevalence in California using Neural Network with confidence interval

2.3 Results

After creating the different models, we make our estimates for the infections and calculate the different fractions. We use the latter to calculate the confidence intervals and then we perform our aggregation so that each column condenses at least the results corresponding to 100 responses by the participants. The purpose of this aggregation is to be able to share the results since according to the data confidentiality contract that binds IMDEA, CMU and Facebook, these data can not be shared by IMDEA unless there is an aggregation of 100 at least. We then create for each state a file corresponding to the aggregated results, for California for example we obtain the table (head, 9) :

	state	date	A5_1	A5_2	A5_3	A2	pos_RF	pos_XGB	pos_glm	pos_svm	pos_nn	count	symptomatic	infected	not_infected	vaccinated
1	California	2021-06-01	2340.0	5554.000	2029.000	1011	30	24	19	-44.82189	62	2509	2410	0	0	2053
2	California	2021-06-02	2370.5	5631.000	2233.072	937	26	22	22	-42.41170	55	2455	2356	0	0	2060
3	California	2021-06-03	2328.0	5556.000	2260.000	968	27	23	21	-44.71228	52	2373	2263	0	0	1938
4	California	2021-06-04	2320.0	5502.100	2132.000	1304	19	15	15	-50.22337	61	2384	2286	0	0	1950
5	California	2021-06-05	1930.0	4950.000	2110.000	960	26	23	22	-39.24137	69	2229	2120	0	0	1844
6	California	2021-06-06	2129.0	4889.000	2064.000	808	20	16	16	-45.50601	52	2256	2154	0	0	1877
7	California	2021-06-07	2090.0	4774.405	1732.000	644	22	21	19	-43.44995	52	2239	2146	0	0	1800
8	California	2021-06-08	1994.0	5105.444	1572.400	821	21	15	16	-48.51323	42	2272	2166	0	0	1876

	unvaccinated	cli	cli_vaccinated	cli_unvaccinated	cli_local	count_local	reach	test_recent	tested_vaccinated	tested_unvaccinated	positive_recent
1	300	23	13	9	157	2498	15904.28	213	172	34	9
2	262	25	19	4	190	2446	15573.20	190	162	25	10
3	279	23	7	11	193	2364	15051.13	202	172	29	7
4	289	19	13	4	283	2376	15127.53	176	147	28	4
5	247	25	14	5	192	2218	14121.57	183	165	17	9
6	252	23	13	6	110	2245	14293.48	194	164	25	5
7	307	24	11	10	173	2232	14210.71	167	142	24	3
8	251	18	7	7	314	2268	14439.91	200	170	27	4

	positive_vaccinated	positive_unvaccinated	positive_symptomatic	days_aggregated	p_positive_recent	p_positive_recent_CI	p_rf	p_rf_CI	p_XGB
1	4	3	9	1	0.04225352	0.02701563	0.011956955	0.004253006	0.009565564
2	7	3	10	1	0.05263158	0.03175075	0.010590631	0.004049219	0.008961303
3	6	1	7	1	0.03465347	0.02522246	0.011378003	0.004267243	0.009692373
4	2	2	4	1	0.02272727	0.02201777	0.007969799	0.003569284	0.006291946
5	8	1	8	1	0.04918033	0.03133050	0.011664424	0.004457352	0.010318528
6	2	2	5	1	0.02577320	0.02229777	0.008865248	0.003868035	0.007092199
7	1	2	3	1	0.01796407	0.02014447	0.009825815	0.004085650	0.009379187
8	3	0	4	1	0.02000000	0.01940265	0.009242958	0.003934894	0.006602113

	p_XGB_CI	p_glm	p_glm_CI	p_nn	p_nn_CI	p_svm	p_svm_CI	p_cli	p_cli_CI	p_cli_local	p_cli_local_CI	p_cli_vaccinated
1	0.003808605	0.007572738	0.003392138	0.02471104	0.006074491	-0.01786444	NA	0.009166999	0.003729165	0.009871558	0.001536490	0.006332197
2	0.003727805	0.008961303	0.003727805	0.02240326	0.005854069	-0.01727564	NA	0.010183299	0.003971403	0.012200444	0.001724175	0.009223301
3	0.003941843	0.008849558	0.003768164	0.02191319	0.005890345	-0.01884209	NA	0.009692373	0.003941843	0.012822961	0.001797444	0.003611971
4	0.003174073	0.006291946	0.003174073	0.02558725	0.006338379	-0.02106685	NA	0.007969799	0.003569284	0.018707617	0.002159099	0.006666667
5	0.004195171	0.009869897	0.004103888	0.03095559	0.007190091	-0.01760492	NA	0.011215792	0.004371785	0.013596219	0.001910042	0.007592191
6	0.003462768	0.007092199	0.003462768	0.02304965	0.006192228	-0.02017111	NA	0.010195035	0.004145221	0.007695819	0.001432613	0.006925946
7	0.003992614	0.008485931	0.003799446	0.02322465	0.006238684	-0.01940596	NA	0.010719071	0.004265397	0.012173919	0.001803000	0.006111111
8	0.003330021	0.007042254	0.003438469	0.01848592	0.005538762	-0.02135265	NA	0.007922535	0.003645430	0.021745284	0.002378890	0.003731343

	p_cli_vaccinated_CI	p_cli_unvaccinated	p_cli_unvaccinated_CI
1	0.003431243	0.04333333	0.02303981
2	0.004128056	0.07251908	0.03140336
3	0.002670900	0.02508961	0.01835166
4	0.003611875	0.04498270	0.02389613
5	0.003961834	0.05668016	0.02883664
6	0.003751857	0.05158730	0.02730976
7	0.003600318	0.03583062	0.02079135
8	0.002759005	0.02788845	0.02036955

FIGURE 2.11 – Table of aggregated results of classification (first 8 rows for California)

2.3.1 Total infections

Now we plot all the curves together to see the correlation, we group together in the same figure (p_cli, p_cli_local, p_glm, p_rf, p_XGB, p_svm-lin, p_nn, p_positive_recent). The values p_positive_recent are the test positivity rate -TPR- observed in the ground truth (*Confirmed* in the figure). We obtain :

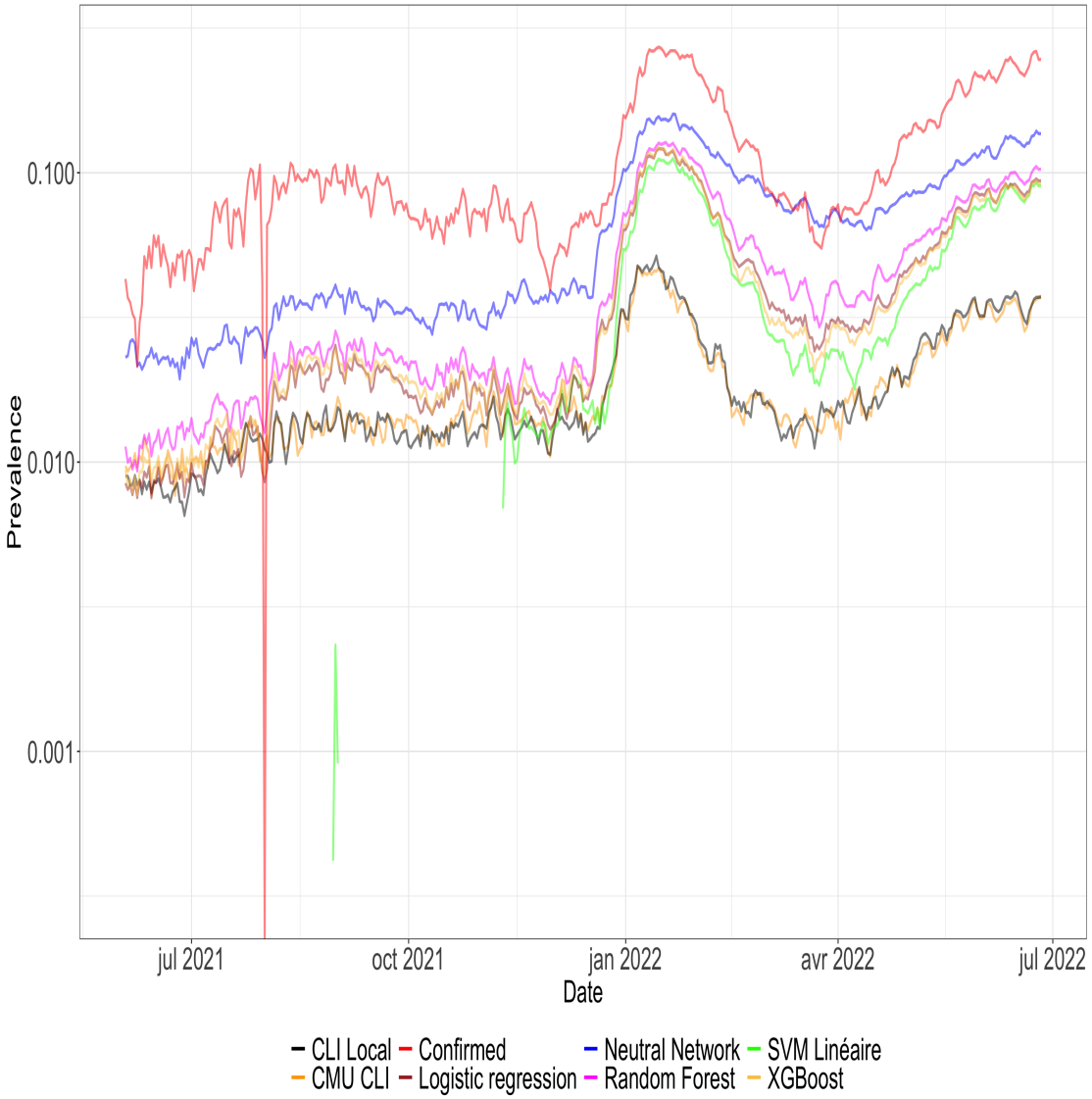

FIGURE 2.12 – Prevalence in California using different classifiers

We generate the following figure giving the correlations to see which models give the most correlated estimate to the TPR of the ground truth (the confirmed

cases) :

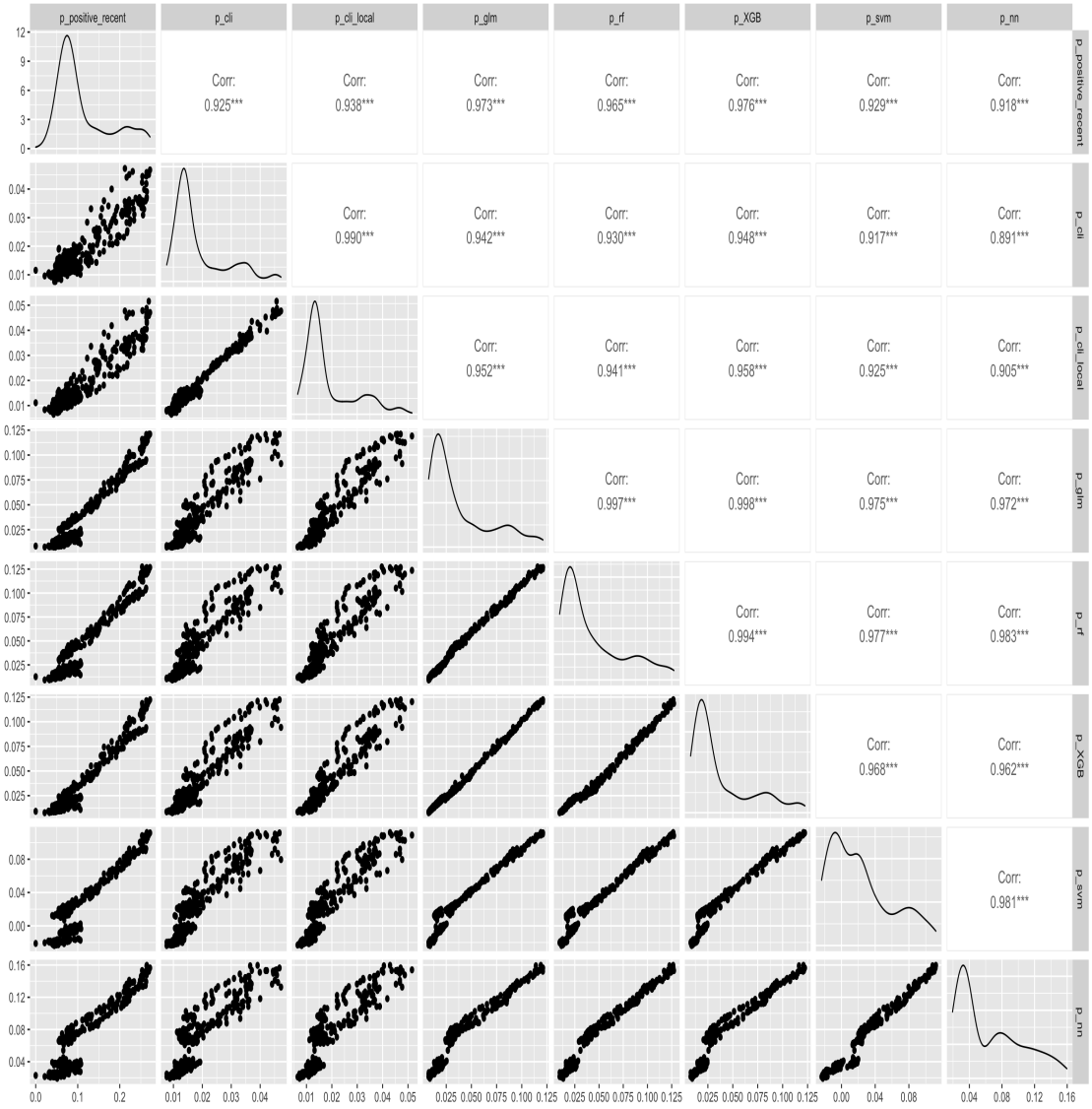

FIGURE 2.13 – Table of correlations

— From the figure of curves, we notice that Neural Network gives the closest approximation to the curve of confirmed cases, while XGBoost, linear SVM, Random Forest and logistic regression respect the correlation more.

— The figure giving the correlations confirms the previous remark.

— We also notice that all the curves have three waves, the first is in the summer of 2021, the second is during the beginning of 2022 and the beginning of the summer of the same year. This can be justified by the release of the containment in the summer of 2021 and the cancellation of mask and barriers recently in 2022, as well as the holidays and gatherings at the beginning of the year to celebrate the new year.

— The estimation corresponding to CLI Local give a good correlation too even if it's based on indirect questions that dont involve directly the participants.

— We also notice that the fraction TPR of confirmed cases is always higher, which can be justified by the fact that people who think they are positive tend to go get tested to confirm whether they have covid or not more than the others.

— The low values of CMU CLI and CLI Local are possibly because the fact that these two methods fix a combinations of symptoms that may not detect well the covid.

2.3.2 Prevalence Vs Vaccination

First, we plot the fractions of people vaccinated between April 2021 and June 2022 :

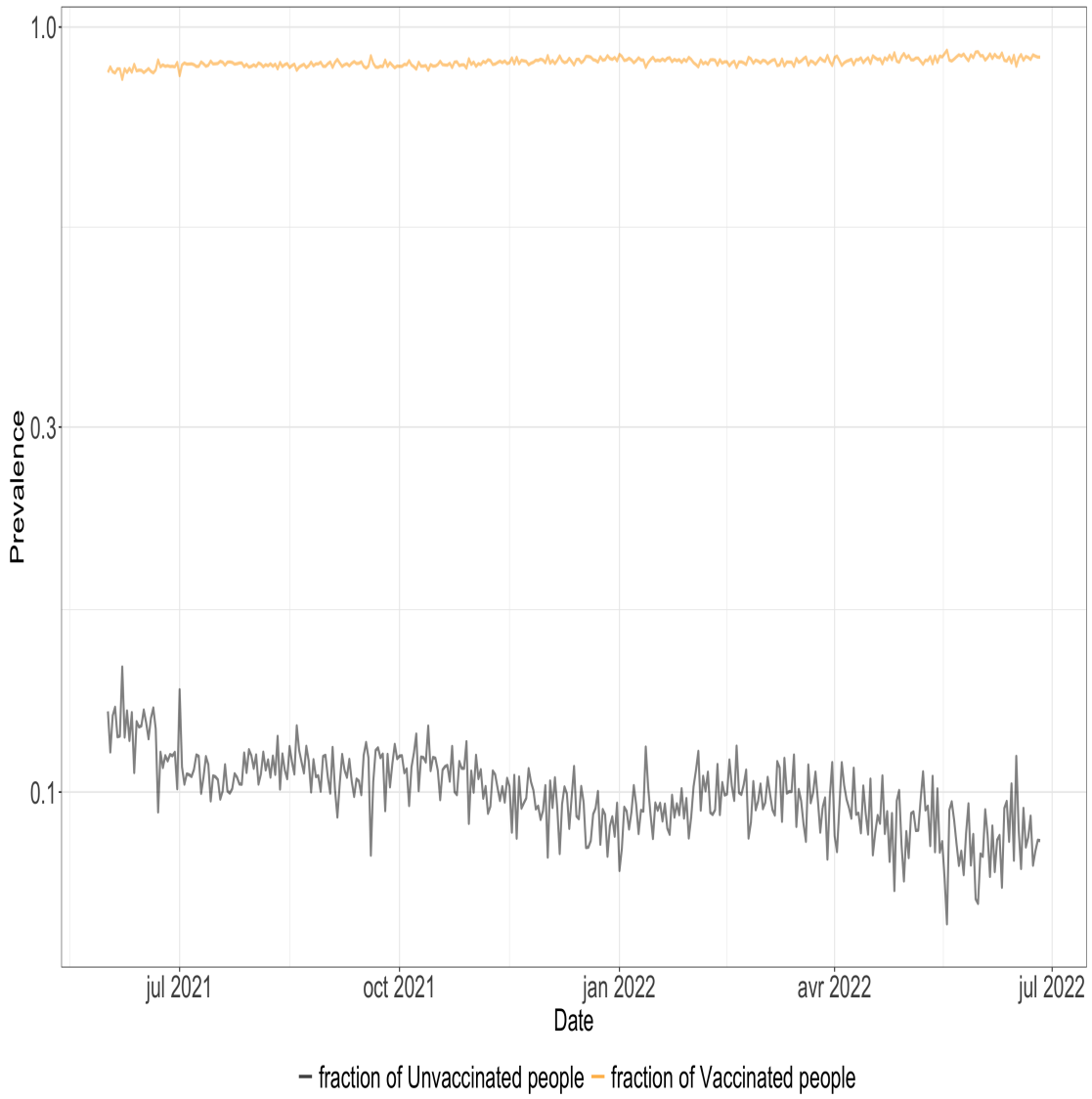

FIGURE 2.14 – Vaccination status in California

The huge difference is justified by the fact that vaccination in the U.S. began early and by June 2021 most of the population was already vaccinated.
In the same figure we plot the fraction of infections among vaccinated people as well as among non-vaccinated people :

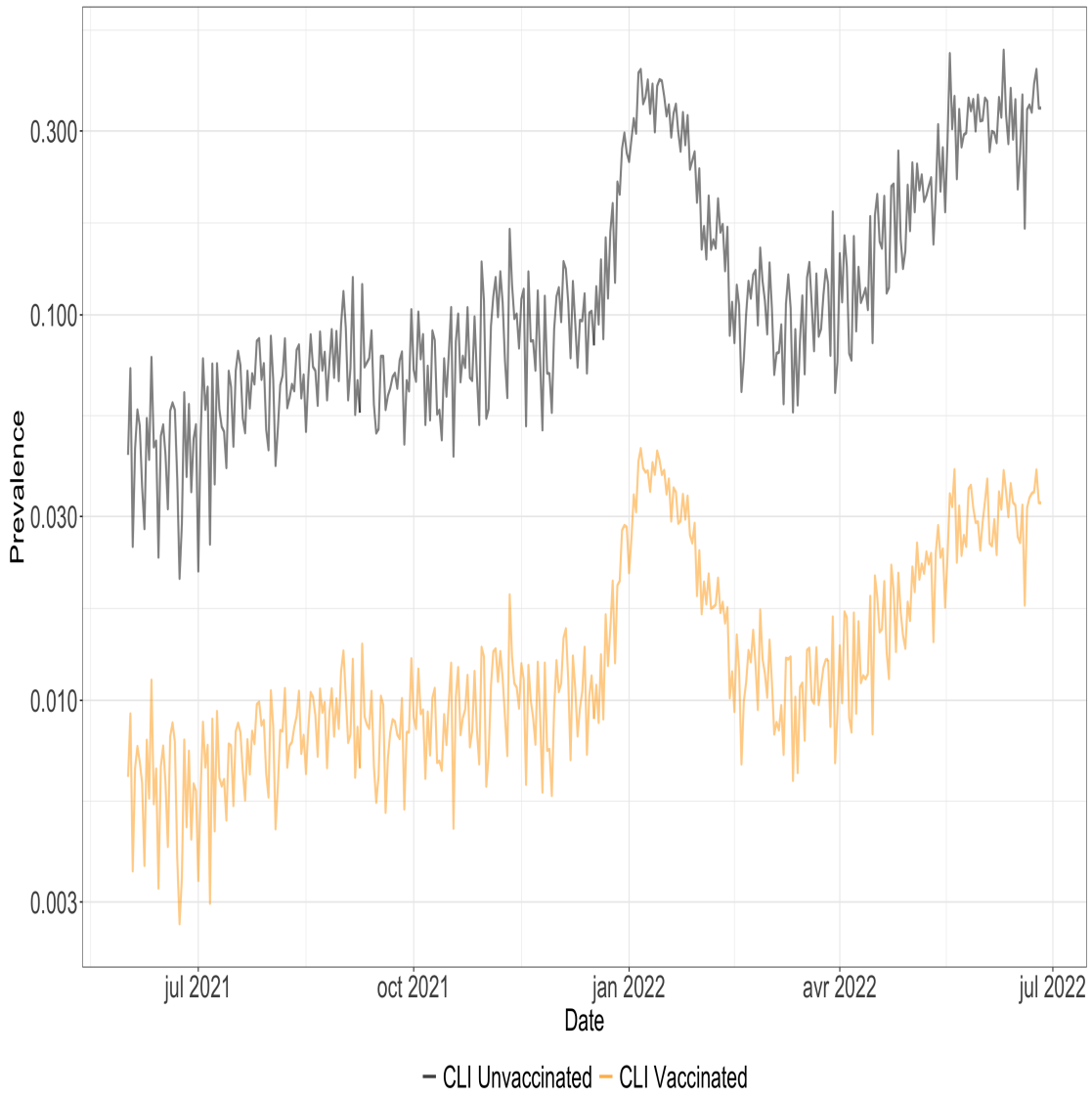

FIGURE 2.15 – Prevalence in California among vaccinated and unvaccinated people using CMU CLI criteria

We note that in California, being unvaccinated the probability of getting covid is almost 10 times higher than being vaccinated.
We also see that both curves are well correlated and have the same seasonality.
Now we plot the prevalence in California among Vaccinated, Unvaccinated with different proxies.

— Vaccinated :

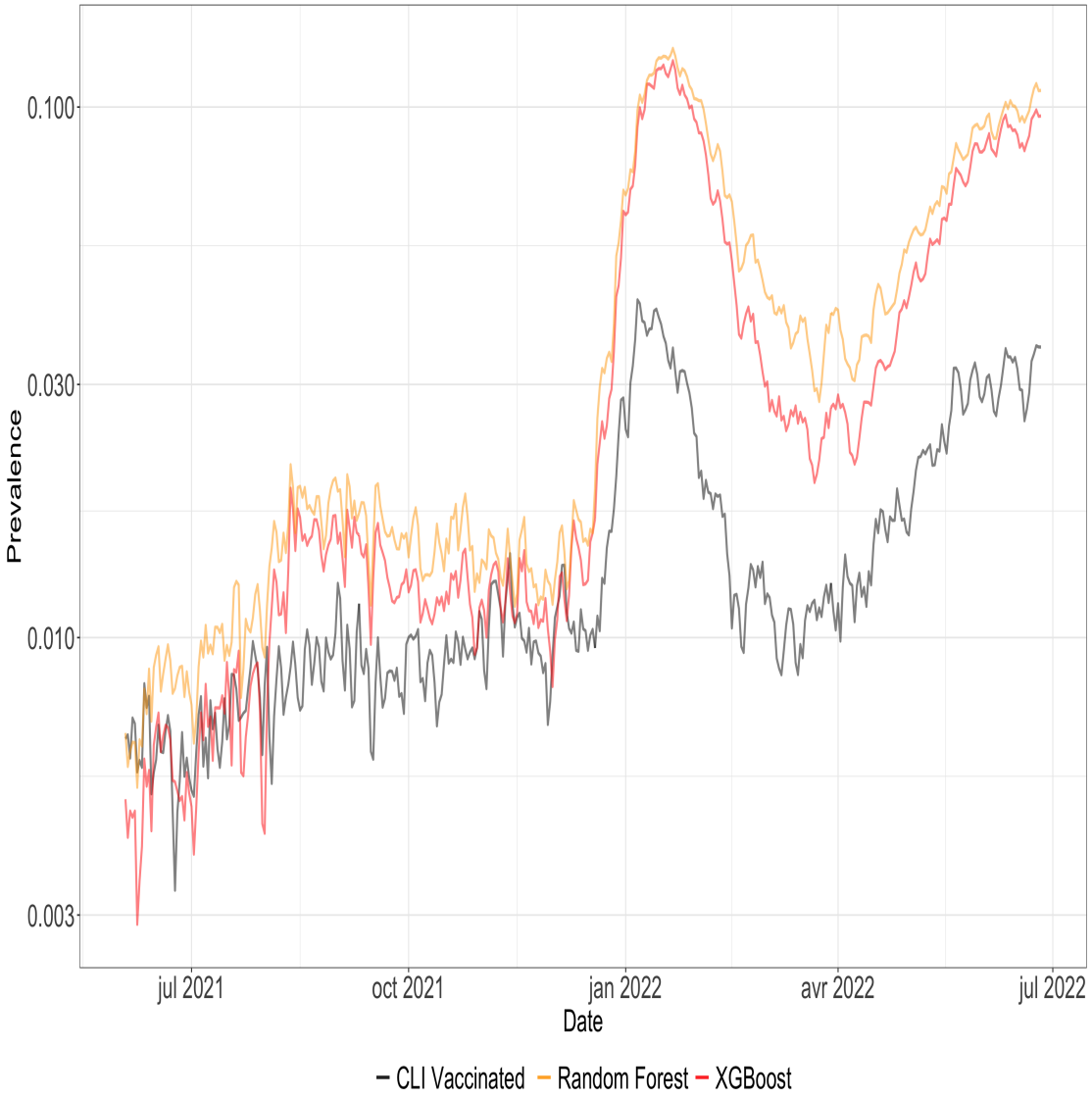

FIGURE 2.16 – Prevalence in California among vaccinated people using CMU CLI, Random Forest and XGBoost

— Unvaccinated :

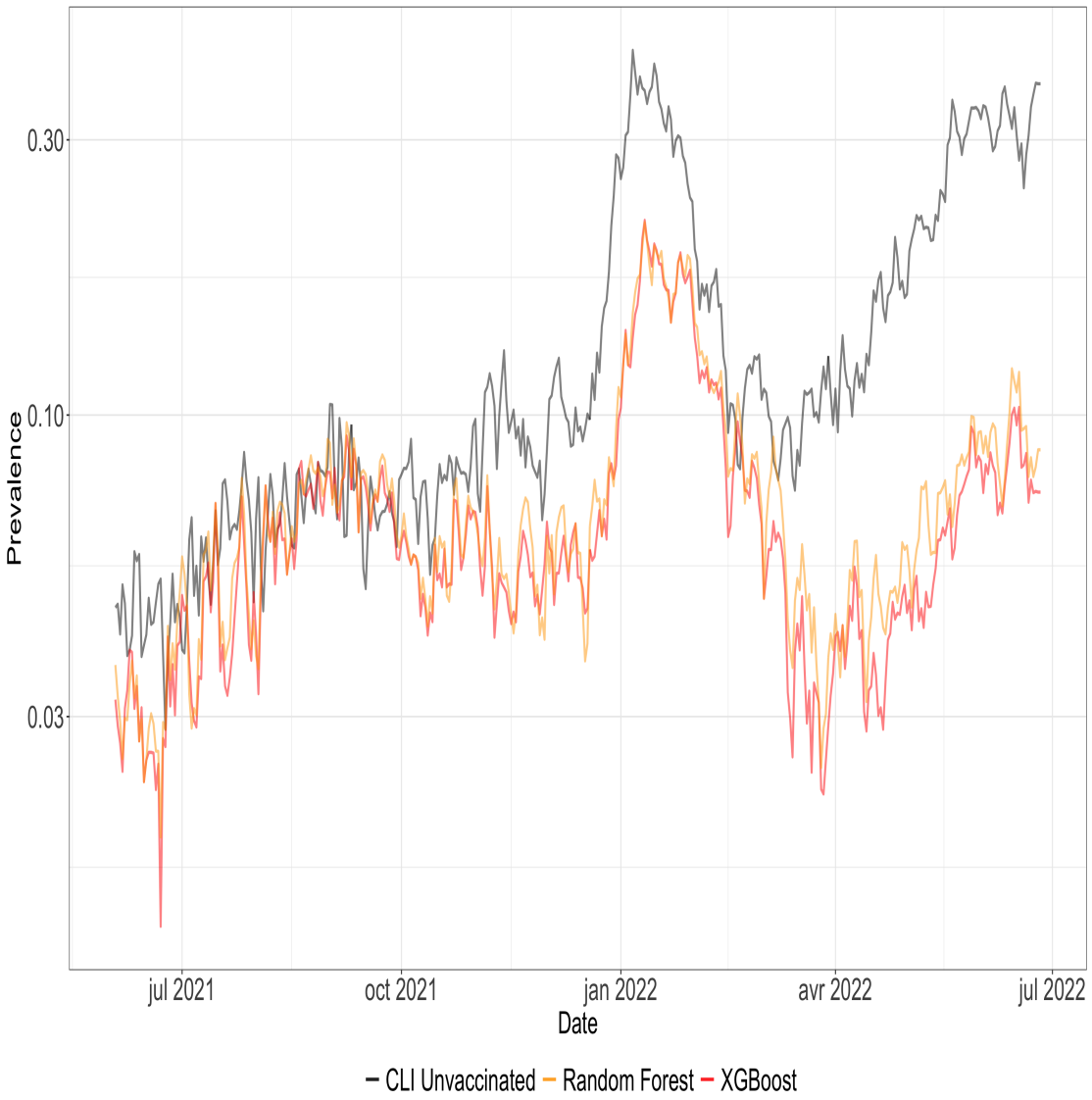

FIGURE 2.17 – Prevalence in California among unvaccinated people using CMU CLI, Random Forest and XGBoost

— first of all, we noticed that Random Forest and XGBoost give higher prevalence for vaccinated people and lower prevalence for unvaccinated. That means vaccination is more detected by CMU CLI. This is possibly because of the fact that CMU CLI is based on symptoms.

— We can conclude that the vaccinated people, when they catch the covid, they have it with less symptoms, so with CMU CLI we detect less positive cases for the vaccinated people.

— Now let's focus on just one of the methods, XGBoost for example. Seeing Vaccinated Vs Unvaccinated prevalence we can see a steeper increase for people vaccinated than for those not vaccinated. And for unvaccinated, the height of the two waves are more close. From that we can conclude that the vaccination effect start to decrease a little bit.

— From the figure of correlation too we can say that we have really good results by the different methods, almost all the correlations are higher than 0.92.

— The low correlation for NN is perhaps due to the fact that we used a random sampling of 5000 to avoid the infinite time that the model generation takes.

Now we plot the prevalence in California among people with different levels of vaccination (vaccinated, unvaccinated, 1st dose, 2nd dose), estimated with Random Forest :

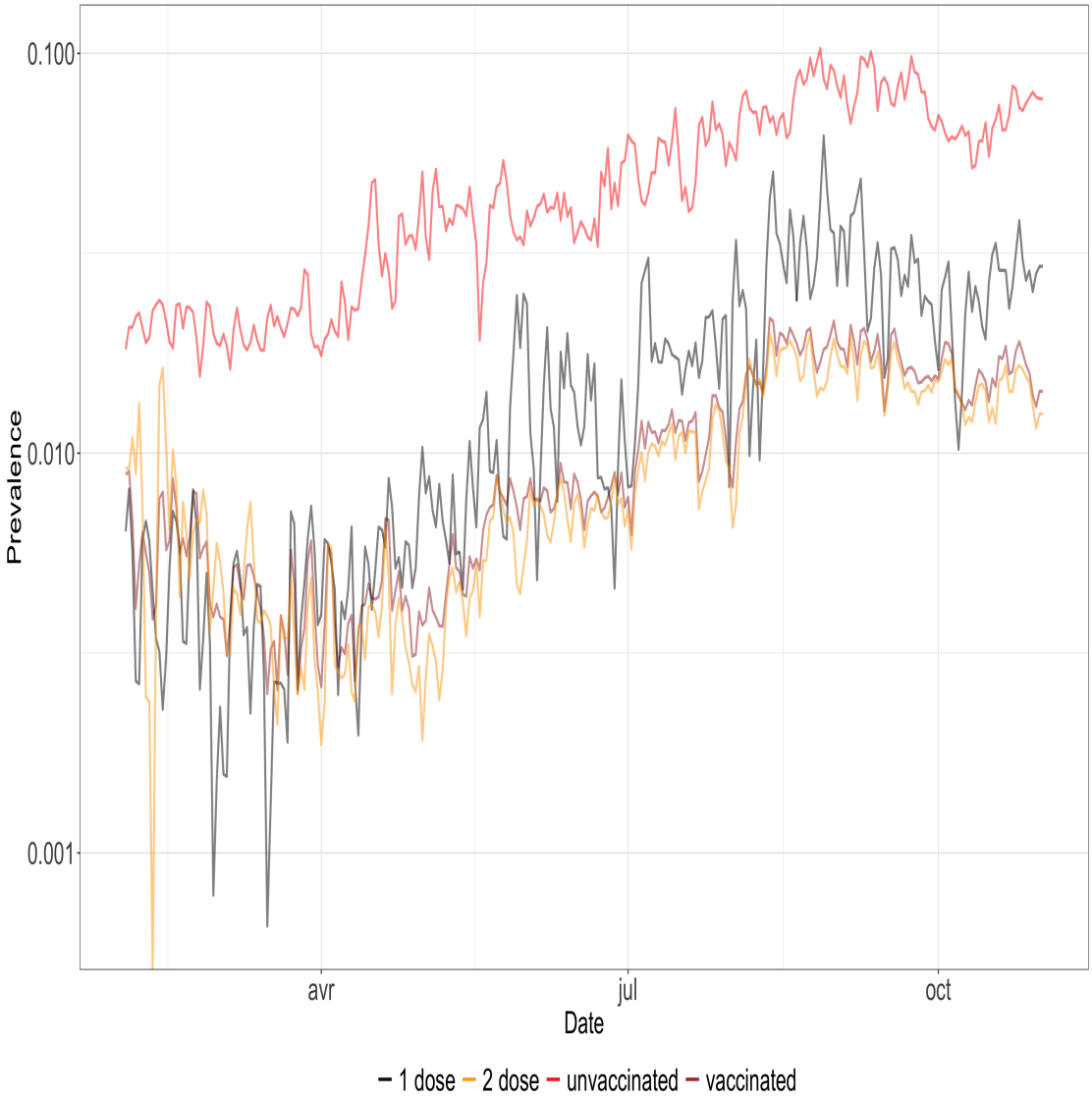

FIGURE 2.18 – Prevalence in California among Vaccinated, Unvaccinated, Vaccinated with 1 dose, and Vaccinated with 2 doses

— We note that unvaccinated people are much more likely to get covid than vaccinated people (almost 10 times more).

— for vaccination we can see that at the beginning of 2021 people having 1 dose or 2 doses have covid with the same fraction, which is possibly because the fact that people just started to get the 2nd dose and it didn't give yet its efficiency.

— at the end of 2021 we can see that the curves of the 1st dose and the 2nd

dose started to be obviously separated because the second dose started to have its effect.

— the curves of Vaccinated and 1st dose are so similar at the end of 2021, which means the most of vaccinated people have done 2 doses.

Finally, we can see from the first curves that include the year 2022 that we have 3 waves, in fact it may be because most people were vaccinated at the beginning, which made the prevalence go down, but after a few months the effect of the first dose started to decrease, so almost all people had reminder doses, so we were able to build up an immunity able to make the curve go down, but we had again a decrease of the effect of the vaccine with the appearance of new variant. This last wave could also be justified by the limitation of sanitary management against Covid recently, but it seems that it will go down soon because the collective immunity has started to be well strengthened.

3 Evaluating the classifiers

3.1 Confusion Matrices - Classification Performance Metrics

To test how good the proxy estimates for each model are, we took verify their performance in different states based on a specific approach. We use a random sampling of the ground-truth and dividing it into a training set (80%) and a testing set (20%).

For our classificaion models, the result is binary (0 : negative, 1 : positive). To judge the performance we do a comparaison between our outcomes and the actual values given in the ground-truth. For that we use a matrix called the confusion Matrix [5].

with :
- TP (True Positive) : the model predicted an outcome of true and the actual observation was true.
- FP (False Positive) : the model predicted a true outcome but the actual observation was false.
- FN (False Negative) : the model predicted a false outcome while the actual observation was true.
- TN (True Negative) : the model predicted an outcome of false, while the actual outcome was also false.

This matrix will help us calculating some performance factors like [5] :
- Accuracy = (TP+TN)/(TP+FP+FN+TN)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1 Score = 2(Precision*Recall)/(Precision+Recall)

F1 score seems to be good for our test because maximising it looks into limiting both false positives and false negatives as much as possible.
For our performance test, in order to be as reliable as possible, we use a repeated cross-validation by taking a different random sample 50 times for each state.
for each state, and for each iteration and sample chosen, we calculate the four performances (Accuracy, Precision, Recall, F1 score) and generate a CSV file containing these values as well as a file containing the probabilities of being positive for each of the models for the given answers, and at the end, we generate a file containing, for each state, the average value of the values of F1 Score of the 50 iterations.

We take the example of some states and we give the averages of the F1 Score that we found for each of the 5 models used :

— California :

```
Average F1 Score model Logistic regression pour California F1 Score =  0.590143
Average F1 Score model XGBoost pour California F1 Score =  0.6106904
Average F1 Score model Random Forest pour California F1 Score =  0.589171
Average F1 Score model Linear SVM pour California F1 Score =  0.5047253
Average F1 Score model Neutral Network pour California F1 Score =  0.5274026
```

FIGURE 3.1 – F1 Score average for the different models - California

— Alabama :

```
Average F1 Score model Logistic regression pour Alabama F1 Score =  0.6764686
Average F1 Score model XGBoost pour Alabama F1 Score =  0.6871228
Average F1 Score model Random Forest pour Alabama F1 Score =  0.6759091
Average F1 Score model Linear SVM pour Alabama F1 Score =  0.5410412
Average F1 Score model Neutral Network pour Alabama F1 Score =  0.634773
```

FIGURE 3.2 – F1 Score average for the different models - Alabama

— Florida :

```
Average F1 Score model Logistic regression pour Florida F1 Score =  0.6143721
Average F1 Score model XGBoost pour Florida F1 Score =  0.6218194
Average F1 Score model Random Forest pour Florida F1 Score =  0.6112979
Average F1 Score model Linear SVM pour Florida F1 Score =  0.4941415
Average F1 Score model Neutral Network pour Florida F1 Score =  0.565651
```

FIGURE 3.3 – F1 Score average for the different models - Florida

— New York :

```
Average F1 Score model Logistic regression pour New York F1 Score =  0.6070842
Average F1 Score model XGBoost pour New York F1 Score =  0.6203828
Average F1 Score model Random Forest pour New York F1 Score =  0.6093708
Average F1 Score model Linear SVM pour New York F1 Score =  0.4937207
Average F1 Score model Neutral Network pour New York F1 Score =  0.5488162
```

FIGURE 3.4 – F1 Score average for the different models - New York

From this values we can conclude that XGBoost has the best performance, followed by Random Forest and Logisitic Regression.
The surprise is that SVM and Neural Network have not a really good performance which was not expected because they are more complicated and take more time. It seems that it's because of the fact that we changed many NA values to 0 in order to avoid some errors for these two models (missing values error) which apparently negatively affected our prediction.
Let's try to confirm these results through the ROC curves.

3.2 ROC curves

The ROC (Receiver Operating Characteristic) curve represents sensitivity versus 1 - specificity for all possible cut-off values of the marker under study. Sensitivity is the ability of the test to detect infected people and specificity is the ability of the test to detect non infected people [5].

```
693  set.seed(12345 + i*7)
694  train_ind<-sample(seq_len(nrow(df_golden)), size = 0.8*nrow(df_golden))
695  train <-df_golden[train_ind,]
696  test <- df_golden[-train_ind,]
```

The Area Under the Curve (AUC) can be interpreted as the probability that, among two randomly selected people, one infected and one not infected, the value

of the marker is higher for the infected than for the not infected. Therefore, an AUC of 0.5 indicates that the marker is non-informative. An increase in the AUC indicates an improvement in discriminatory abilities, with a maximum of 1.0 [5]˙
First, to perform the comparison via the ROC curve in a reliable and fair manner, we randomly sample 5000 of the ground-truth for each state and repeat this 50 times. For each state we then make a loop for i from 1 to 50 :

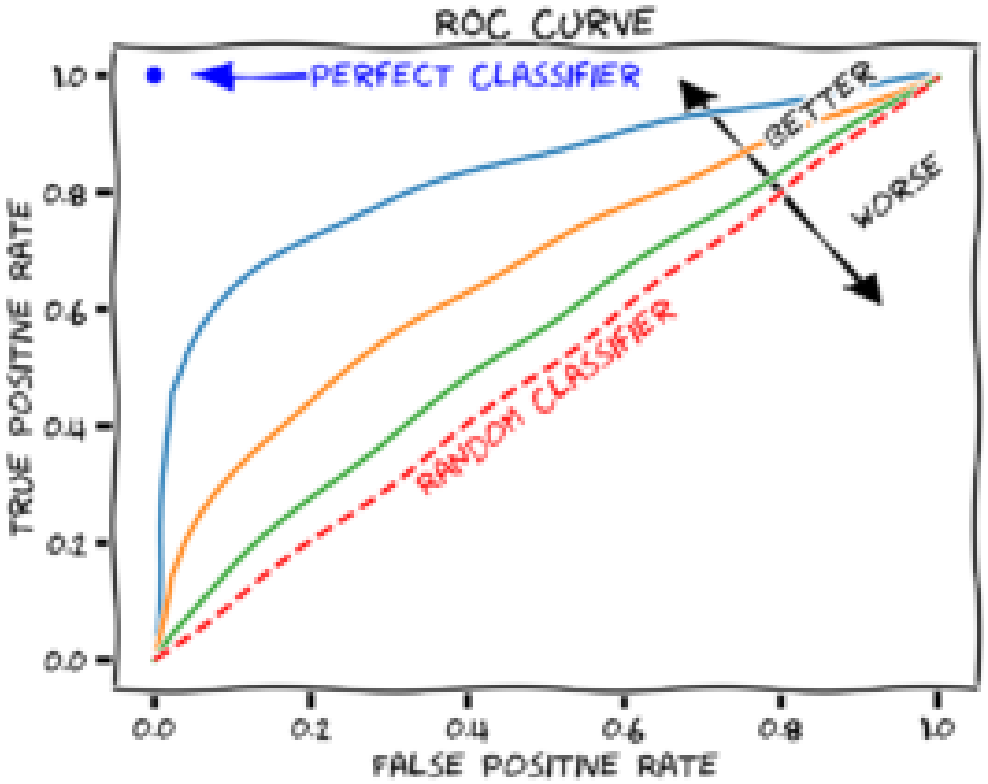

FIGURE 3.5 – reading ROC curves [5]

Then, for each of these samples, we generate a file containing the estimates of the positive cases for the different models as well as the probabilities that will help us plot the ROC curves.
Finally we combine the 50 files to have a file of 50000 (50*(5000*20%)) lines for each state containing the probabilities of being positive for each model.
Thus, we plot the ROC curves of all the models in the same figure :

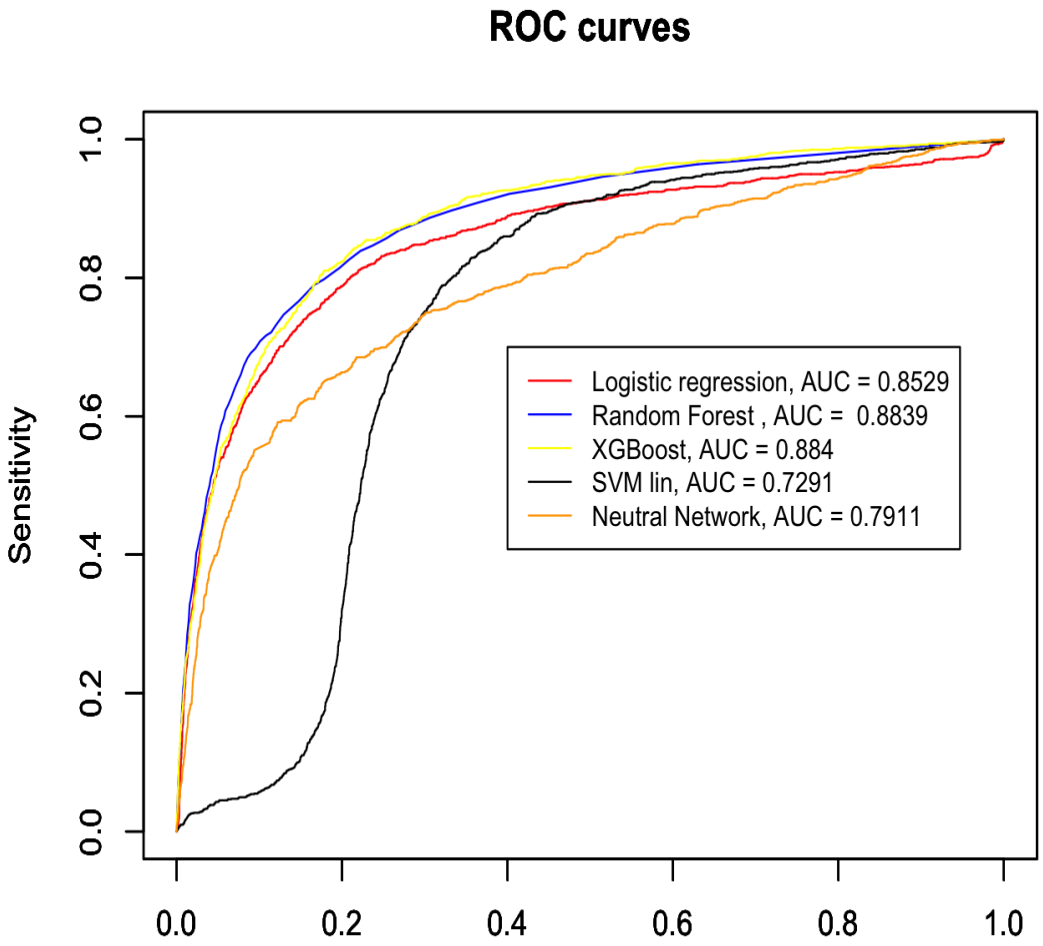

FIGURE 3.6 – ROC curves for logistic regression, Random Forest, XGBoost, Linear SVM and Neural Network

These curves allow us to confirm the previous conclusions concerning the performances of the different models. We notice that the NN and the linear SVM give worse performance, even compared to the logistic regression, which was not expected because these two models are very well known to have very good performances in terms of accuracy. As before, we note a very good performance of XGBoost and Random Forest, which justifies the use of the latter to test the effectiveness of vaccination. The AUC values reinforce the results that appear from

the curves and show that XGBoost and Random Forest were the precise and gave the most accurate results.

Conclusion

This project allowed me to work on real and confidential databases on a subject that concerns the whole world and influences all humanity. It allowed me to broaden my knowledge on the pandemic and on the virus with its different variants. I was also able to familiarize myself with several data science techniques, as well as with the different algorithms and machine learning models and the different parameters that characterize each of them.
This project also allowed me to get initiated to the field of research by starting from a set of data and an idea, which is the study of the pandemic, and generate huge results that characterize the different aspects of the pandemic, its evolution during the last months and the effect of vaccination on this evolution by trying to find explanations of these results, which is not always perfectly correct or definitely scientific explanations but this part of making hypotheses and trying to develop it is an integral part of research and scientific methodology that allow us to better understand this evolution and its continuation in the future.

Discussion

Our results indicate the descent of the efficacy of the SARS-Cov2 vaccine with time, this descent may be due to the variation of the virus is the appearance of new variant with more developed immunity against this vaccination. A recent study has shown for example that vaccination against symptomatic disease is significantly lower against the Omicron variant than with the Delta variant.

It is true that the large number of people who are tested in the United States and the distribution of vaccine throughout the United States has made the study a little easier to detect positive cases or to evaluate the effect of vaccination. However, in order to use these results for other regions of the world or to make a similar approach, one must take into account the large difference in population and the availability of the vaccine, since the United States received the vaccine on a priority basis.

Our study shows that vaccination has a considerably good effect in protecting against this pandemic, but it is not permanent over time, and this leads to ask the question. Do we need regular booster doses against this virus ? This question hides a big problem between people who prefer to avoid vaccination and adapt the collective life to live with the virus until its disappearance, and others who see in vaccination our only rampart in the fight of humanity against Covid-19.

Ethical Declaration

Microdata cannot be made publicly available for ethical and legal reasons, i.e. public availability would compromise confidentiality as data tables list single counts of individuals rather than aggregated. Aggregated data can be made public."

The Ethics Board of IMDEA Networks Institute gave ethical approval to use the data for this work on May 2022 by my arrival and I have the right to use it only as an IMDEA member.

IMDEA Networks has signed Data Use Agreements with Facebook, Carnegie Mellon University (CMU) and the University of Maryland (UMD) to access their data, specifically UMD project 1587016-3 entitled C-SPEC : Symptom Survey : COVID-19 and CMU project STUDY2020 00000162 entitled ILI Community-Surveillance Study.

Bibliography

[1]CoronaSurveys Team. CoronaSurveys : Independent Measurement of the Pandemic through Open Surveys. https ://coronasurveys.org

[2] Effectiveness of COVID-19 vaccines against the Omicron (B.1.1.529) variant of concern.
https ://www.medrxiv.org/content/10.1101/2021.12.14.21267615v1

[3] Using Survey Data to Estimate the Impact of the Omicron Variant on Vaccine Efficacy against COVID-19 Infection.
https ://www.medrxiv.org/content/10.1101/2021.12.09.21267355v1.full

[4] Estimating Active Cases of COVID-19
https ://www.medrxiv.org/content/10.1101/2021.12.09.21267355v1.full

[5] Guide to Confusion Matrices & Classification Performance Metrics.
https ://towardsdatascience.com/guide-to-confusion-matrices-classification-performance-metrics-a0ebfc08408e

[6] Confidence intervals for ratios of proportions : implications for selection ratios.
https ://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12304

[7] Effectiveness of COVID-19 vaccines against the Omicron (B.1.1.529) variant of concern https ://www.medrxiv.org/content/10.1101/2021.12.14.21267615v1.full-text

[8] Machine Learning : Performance and interpretability.
https ://datascientest.com/performance-and-interpretability-in-machine-learning

[9]How to read a ROC curve https ://www.idbc.fr/tutoriel-comment-lire-une-courbe-roc-et-interpreter-son-auc/

[10] neuralnet : Train and Test Neural Networks Using R.
https ://datascienceplus.com/neuralnet-train-and-test-neural-networks-using-r/

[11] Random Forest : Decision Tree Forest - Definition and operation.
https ://datascientest.com/random-forest-definition

[12] La régression logistique, qu'est-ce que c'est ? https ://datascientest.com/regression-logistique-quest-ce-que-cest

[13] Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost. https ://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost