



TADA

PROJECTS

SOLUTIONS

PROJECTS

Let's show your projects

FAQ

HELP



New Project

Scope	Sample Count	ACC	TPR / Sensitivity	TNR	PPV / Precision	F1 Score	MCC
Training	2,410	0.9975	0.9957	0.9982	0.9974	0.9967	0.9968
Validation	1,007	0.9972	0.9970	0.9974	0.9974	0.9972	0.9970
Test	1,204	0.9975	0.9970	0.9982	0.9975	0.9972	0.9973

LET'S GET STARTED!

The project list shows the full list of projects that have been created by the user.

TADA workflow is divided into 5 steps:

1. Project - Select a data source and project
2. Goal - Select the target variable
3. Variable set - Select the variables to be used by the model
4. Model - Select the model to be used
5. Run - Run the model

True positive False positive



True negative False negative



	Predicted Positive	Predicted Negative
Actual Positive	166	1
Actual Negative	2	143
Total	168	144



MYDATAMODELS

THE SMALL DATA PREDICTIVE MODELING COMPANY



THE SMALL DATA
PREDICTIVE MODELING
C O M P A N Y

MyDataModels, la seule
solution métier de
modélisation prédictive
pour traiter les Small Data.



TADA: L'INTERPRÉTATION DES MODÈLES DE PRÉDICTION COVID-19

Jonathan DAEDEN
29-06-2020



1. Le contexte

➔ Dr. David Darmon, Vice-Président politique de santé à Université Côte d'azur (UCA) et Directeur du département d'enseignement et de recherche en médecine générale à la faculté de médecine de Nice

+

➔ Direction des systèmes d'information de l'Université

=

Une plateforme collaborative qui permet aux médecins généralistes de mieux diagnostiquer les cas de patients susceptibles d'être atteints du covid-19.

L'université de Nice lance une plateforme de recherche collaborative sur le CoVid-19

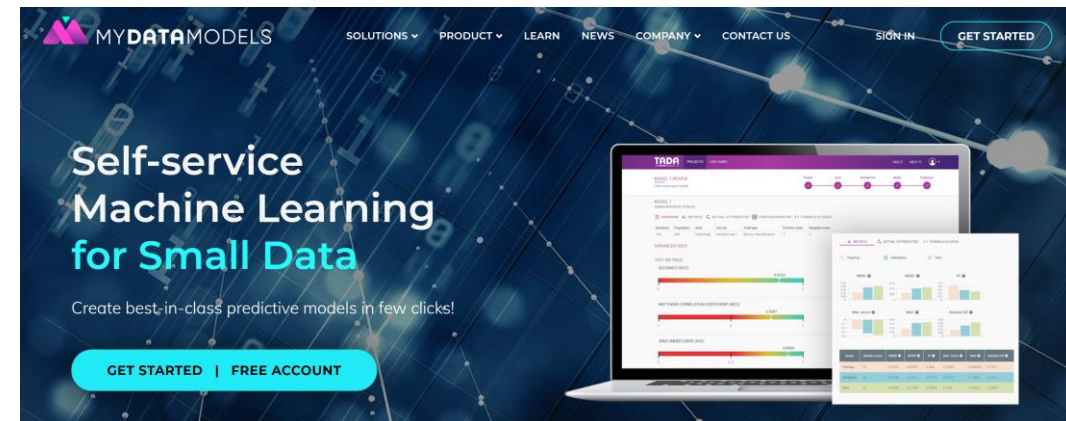
 #NICE  #SANTÉ | PAR E.G. | Mis à jour le 24/03/2020 à 16:59 | Publié le 24/03/2020 à 16:55

Covid-19 : UCA ouvre une plateforme collaborative pour aider les médecins généralistes

Posté ven 27/03/2020 - 08:12 Par [Jean-Pierre Largillet](#)

- 957 médecins généralistes
 - 16 régions françaises
- Plus de 1800 consultations (Mai 2020)

- ➔ MyDataModels, à Sophia Antipolis, développe une solution logicielle de machine learning.
- ➔ Plateforme IA gratuite pour les organisations publiques dans la lutte COVID-19
- ➔ Création d'algorithme IA orienté Small data et interprétable
- ➔ Accompagné par Marc Schoenauer, Directeur scientifique adjoint INRIA + 5 chercheurs PhD spécialisés Machine Learning. 30 personnes
- ➔ Lauréate de la 21ème édition du concours i-Lab by MESRI - 2019
Lauréate Innovative SME Digital South by Pôle SCS – 2018
Accélérée by Allianz, LVMH et THALES.

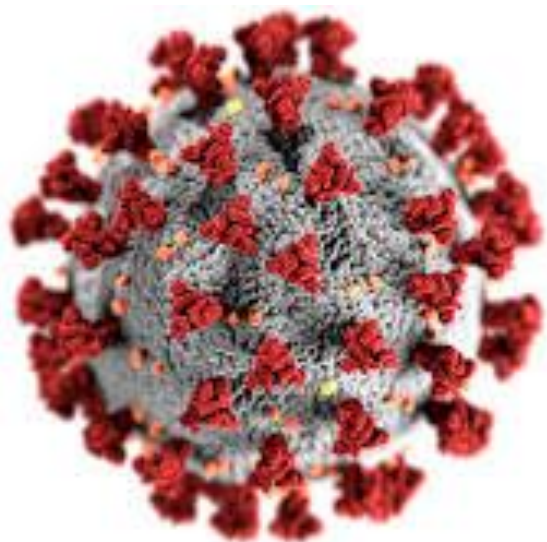


ANNOUNCEMENTS, DATA SCIENCE, PRESS, PRODUCTS

MyDataModels offers Free Access to its Self-Service Machine Learning Platform for Public Organizations fighting COVID-19.

► Learn more

<https://www.mydatamodels.com/>



➡ Problématique COVID-19:

- Médecin généraliste en première ligne face à cette nouvelle maladie
- SARS-CoV-2 encore largement méconnue (symptômes, effet long terme, séquelles)
- Diagnostic et prise en charge des patients suspectés COVID difficile
- Tests systématiques encore impossible.

- ➡ Collaboration sur la problématique suivante:
- Aider la médecine de ville à mieux diagnostiquer efficacement le COVID-19
 - Créer un modèle qui facilite l'identification des facteurs principaux lié au COVID-19
 - Développer un outil d'aide à la décision pour la médecine de ville face au COVID-19



2. La base de données issus de la plateforme de santé

➡ Première étude POC sur un jeu de données réduit:

- Uniquement des patients issus du 06
- 117 patients différents (avec suivi ou non)
- Pour un total de 144 consultations
- Une majorité de diagnostic COVID-19.

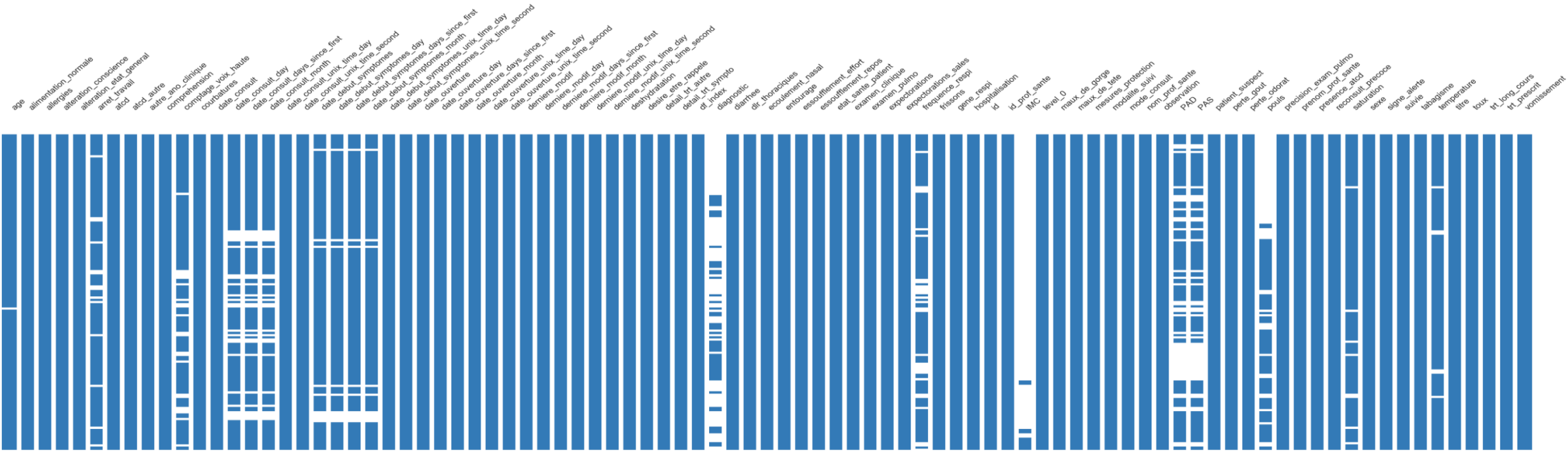
Dataset info

Number of variables	89
Number of observations	144
Missing cells	572 (4.5%)
Duplicate rows	0 (0.0%)

Profilage effectué sous Python 3.6 avec la librairie Pandas Profiling sur un jeu de données anonymisées.

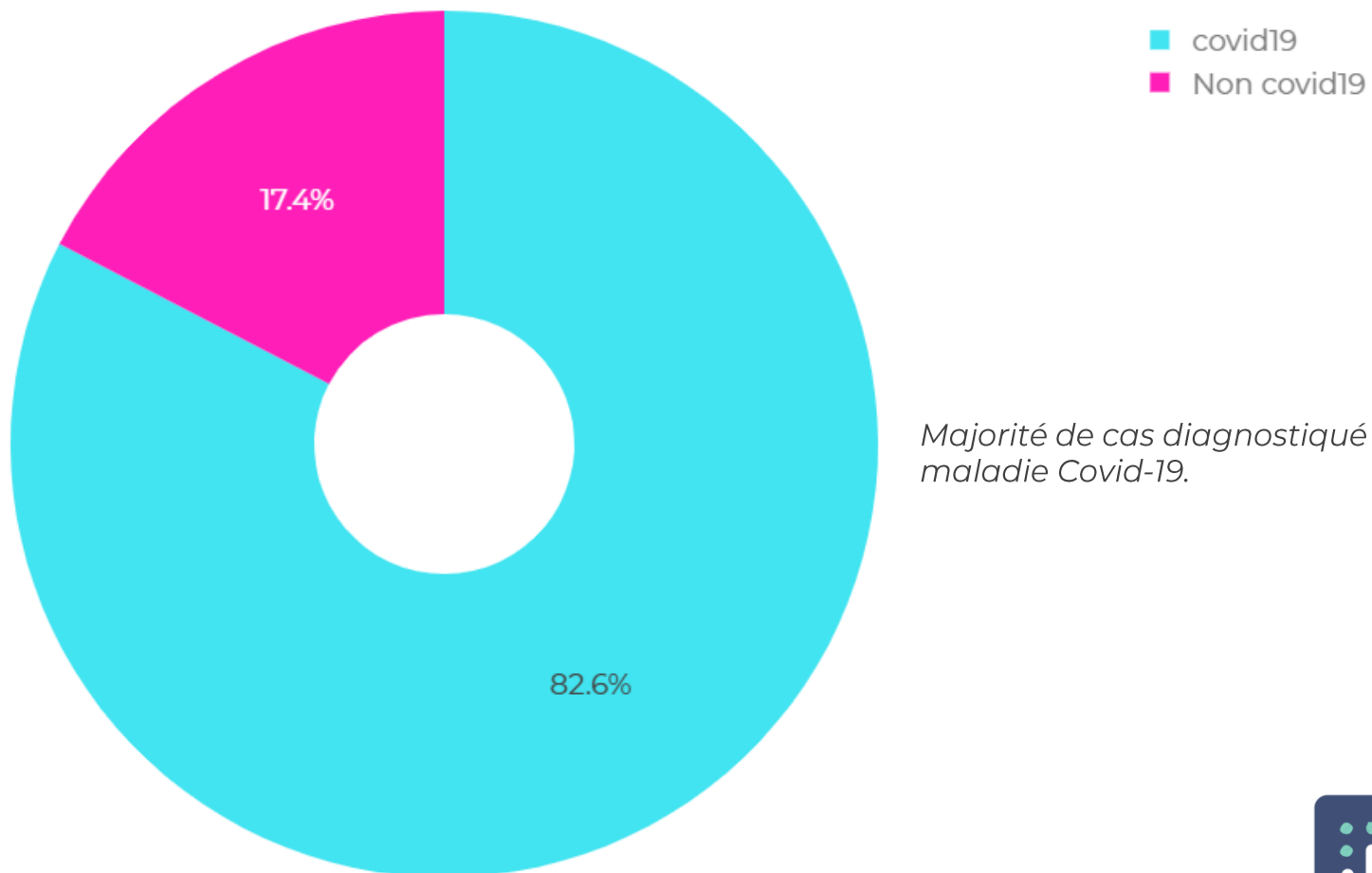


1



Matrice du jeu de donnée tabulaire. Les lignes blanches sont des données manquantes.

Distribution de diagnostic covid19



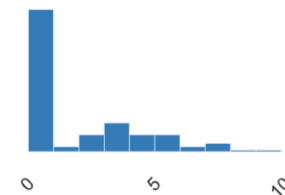
Majorité de cas diagnostiqué atteint de la maladie Covid-19.

➡ Exemple de symptômes suivis lors des consultations (pour un total de 60):

dlr_thoraciques
Numeric

Distinct count 10
Unique (%) 6.9%
Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0

Mean 1.631944444
Minimum 0
Maximum 10
Zeros (%) 58.3%



ecoulement_nasal
Categorical

Distinct count 2
Unique (%) 1.4%
Missing (%) 0.0%
Missing (n) 0

[Toggle details](#)

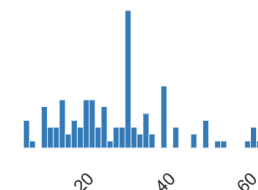
Common values [Composition](#)

Value	Count	Frequency (%)
non	95	66.0%
oui	49	34.0%

comptage_voix_haute
Numeric

Distinct count 41
Unique (%) 28.5%
Missing (%) 15.3%
Missing (n) 22
Infinite (%) 0.0%
Infinite (n) 0

Mean 28.5
Minimum 5
Maximum 65
Zeros (%) 0.0%



➡ Notre objectif:

- Trouver au sein des 60 paramètres suivies lors des consultations (présentielle ou non), les symptômes les plus discriminants (précurseur ou non)
- Prédire la probabilité d'un patient en consultation d'avoir contracté le COVID-19
- Expliquer pourquoi notre modèle estime que le patient a le COVID
- Aider à une meilleure prise en charge du patient



3. Notre méthodologie

➡ 1 / Nettoyage du jeu de données

- Erreur de typographie (température de 383 °C)
- Encodage incorrect ("guÈri")
- Duplication de ligne
- Remplacement des données manquantes IMC has 134 (93.1%) missing values
- Pivot sur les consultations et le suivie.
- Suppression manuelle des colonnes ou lignes erronées/incomplètes (ligne de test remplie de zéro)

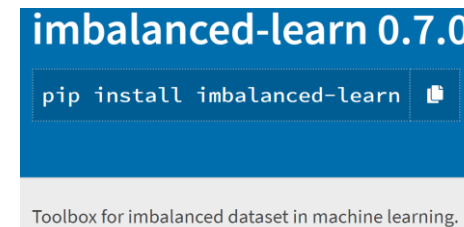
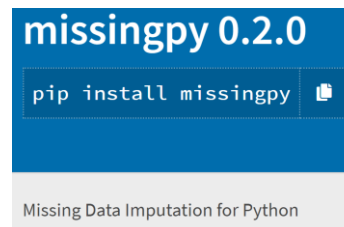


➡ 2 / Feature engineering

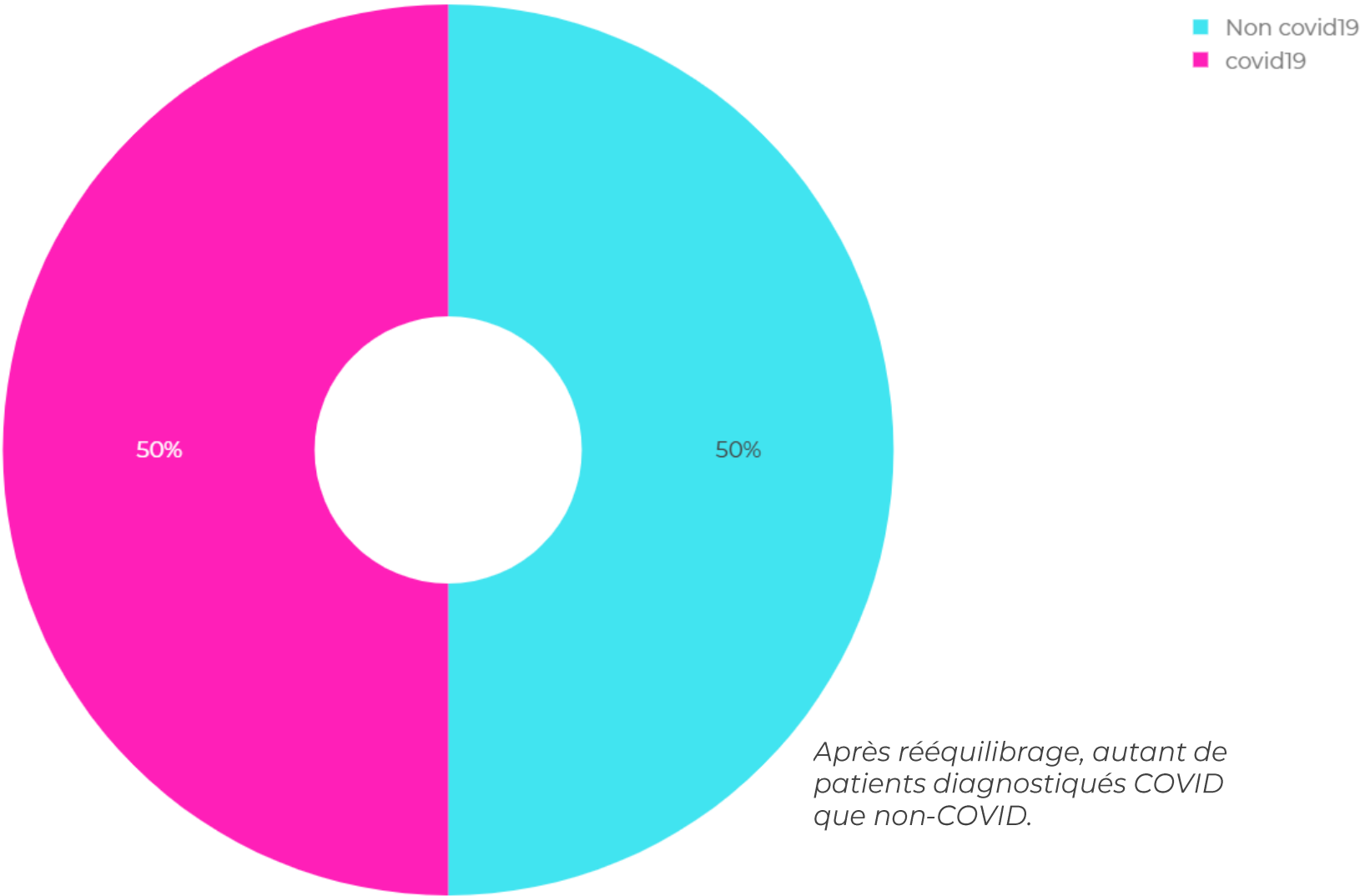
- Imputation de données manquantes avec la librairie missingpy
<https://pypi.org/project/missingpy/>
- Rééquilibrage de l'apprentissage avec imbalanced-learn
<https://pypi.org/project/imbalanced-learn/>
- Encodage et sélection des paramètres primordiaux avec scikit-learn
<https://pypi.org/project/scikit-learn/>

	NaN_percentage
Descriptors	
pouls	43.055556
PAS	31.250000
PAD	31.250000
frequence_respi	20.138889
comptage_voix_haute	15.277778
saturation	5.555556
temperature	4.166667
age	0.694444

Pourcentage de données manquantes dans les paramètres



Distribution de diagnostic covid19





4. La modélisation avec

TADA
ÉDITIONS



TECHNOLOGIE UNIQUE

- Moteur mathématique unique
- Combine les approches symboliques et numériques
- Inspiré des algorithmes évolutionnaires et génétiques
- Opérationnel et performant avec peu de données
- Partenariat scientifique avec INRIA Sophia Antipolis / Université Cote d'Azur

PRODUIT DIFFERENTIANT

- Simple d'usage et d'implémentation
- IA interprétable & compréhensible
- Aide à l'analyse et décisions
- Rapide, compact et sécurisé
- Confidentialité des données – HDS.



TADA® rend l'IA simple & accessible.

Les modèles de ML peuvent être caractérisées par deux critères:

Performance et interprétabilité

Performance : Mesure de la qualité des prédictions, groupes, ou du comportement.

Interprétabilité : Mesure de la simplicité et de la confiance que l'on peut accorder aux modèles.

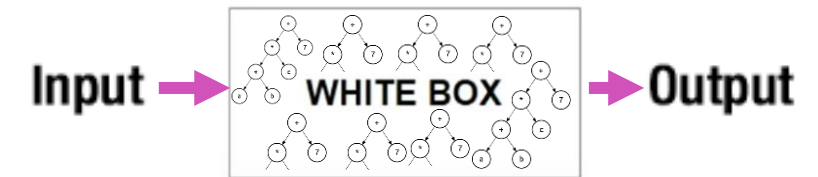
Ces deux critères sont souvent antagonistes :

On peut créer un modèle complexe et performant mais rarement un modèle plus simple et plus performant.

Chez MDM, on cherche à obtenir les modèles les plus interprétable avec la plus haute performance possible.



On s'interdit l'utilisation des algorithmes dit « black box »

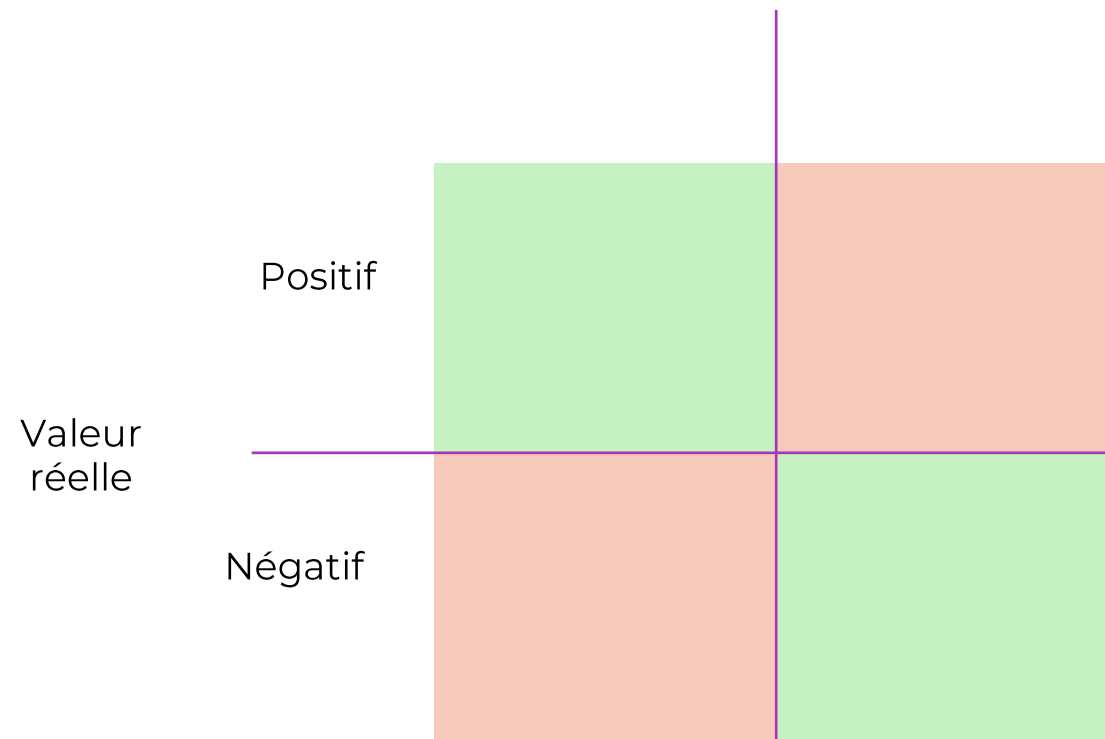


Formules en clair et raisons des prédictions

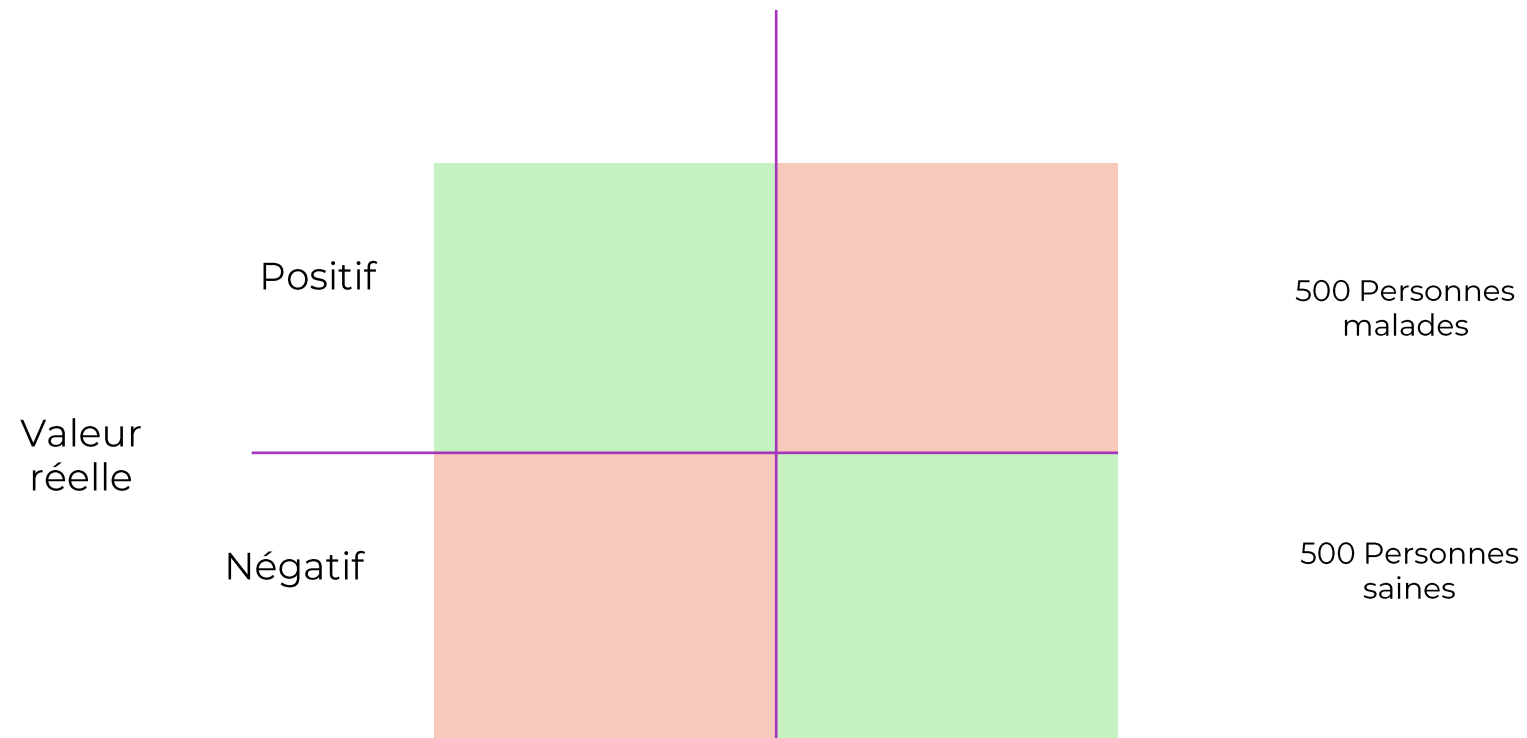


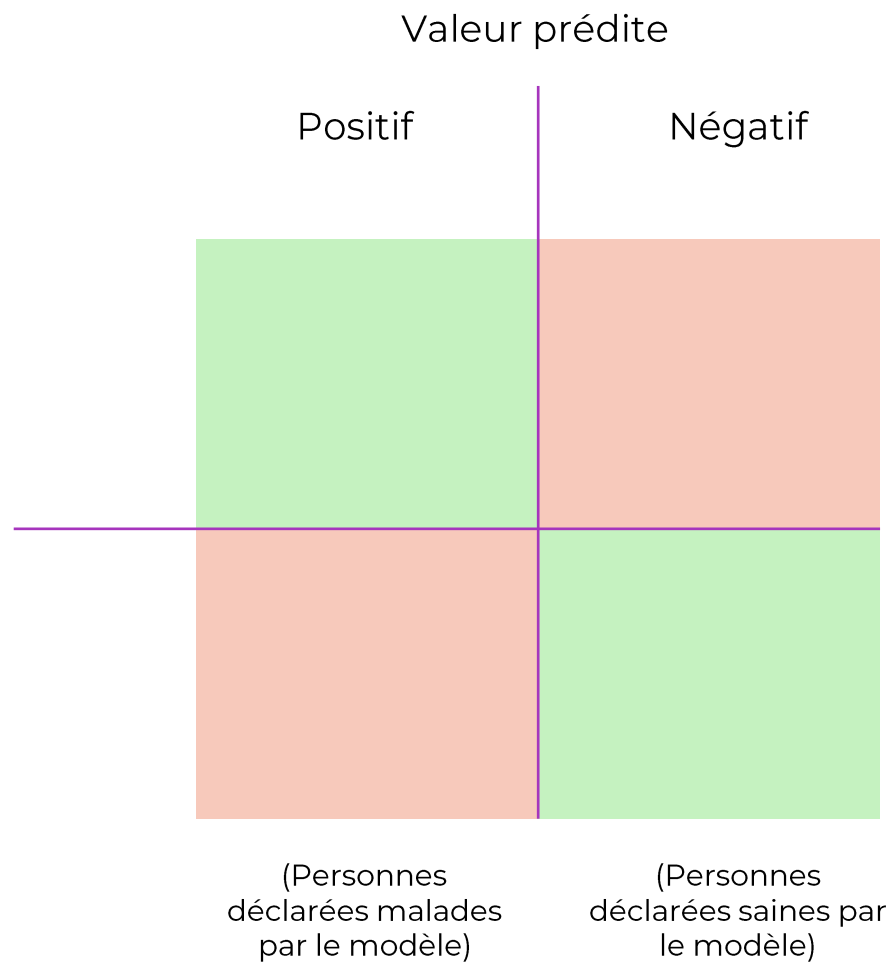


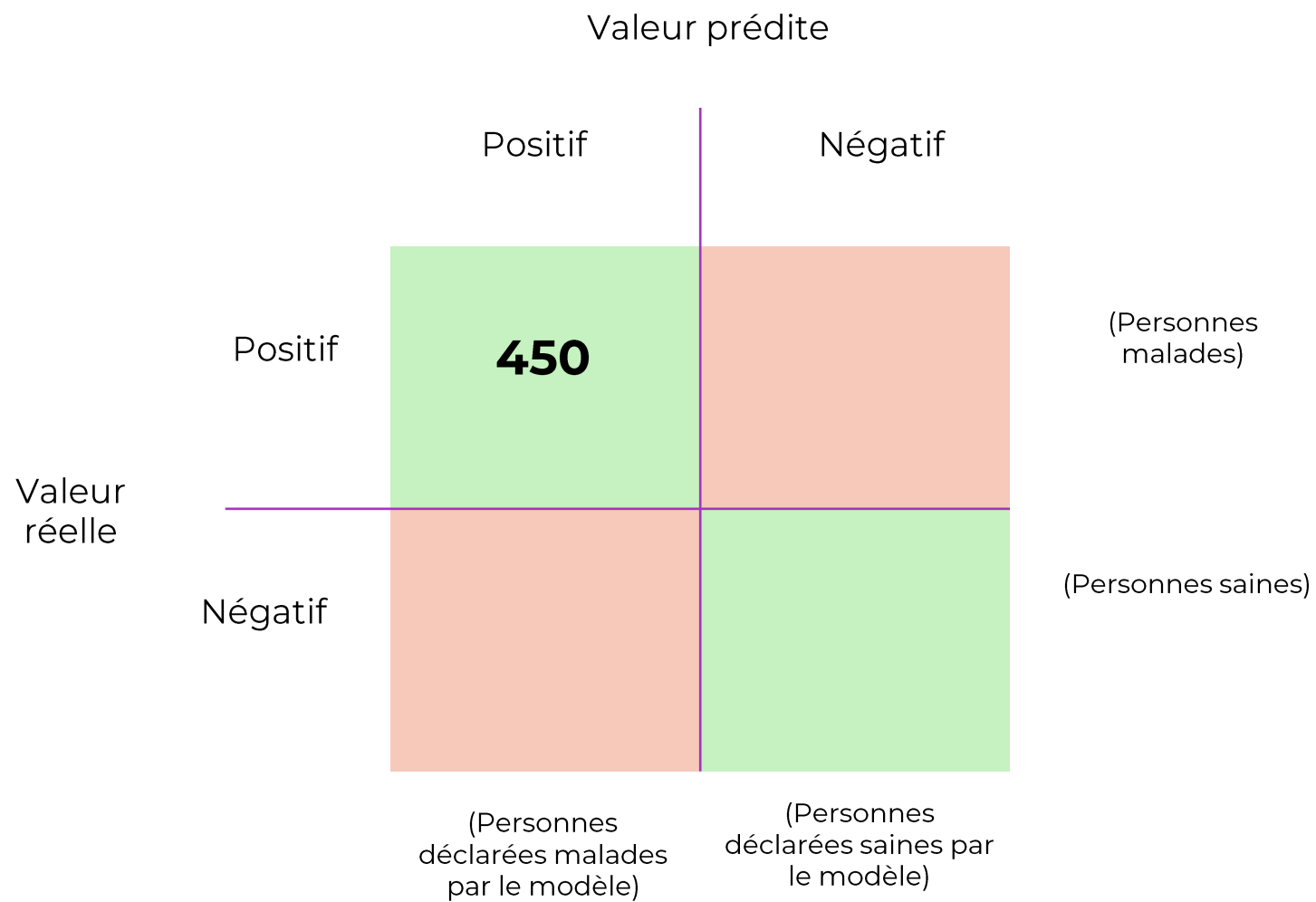
5. Mesurer la performance de nos modèles

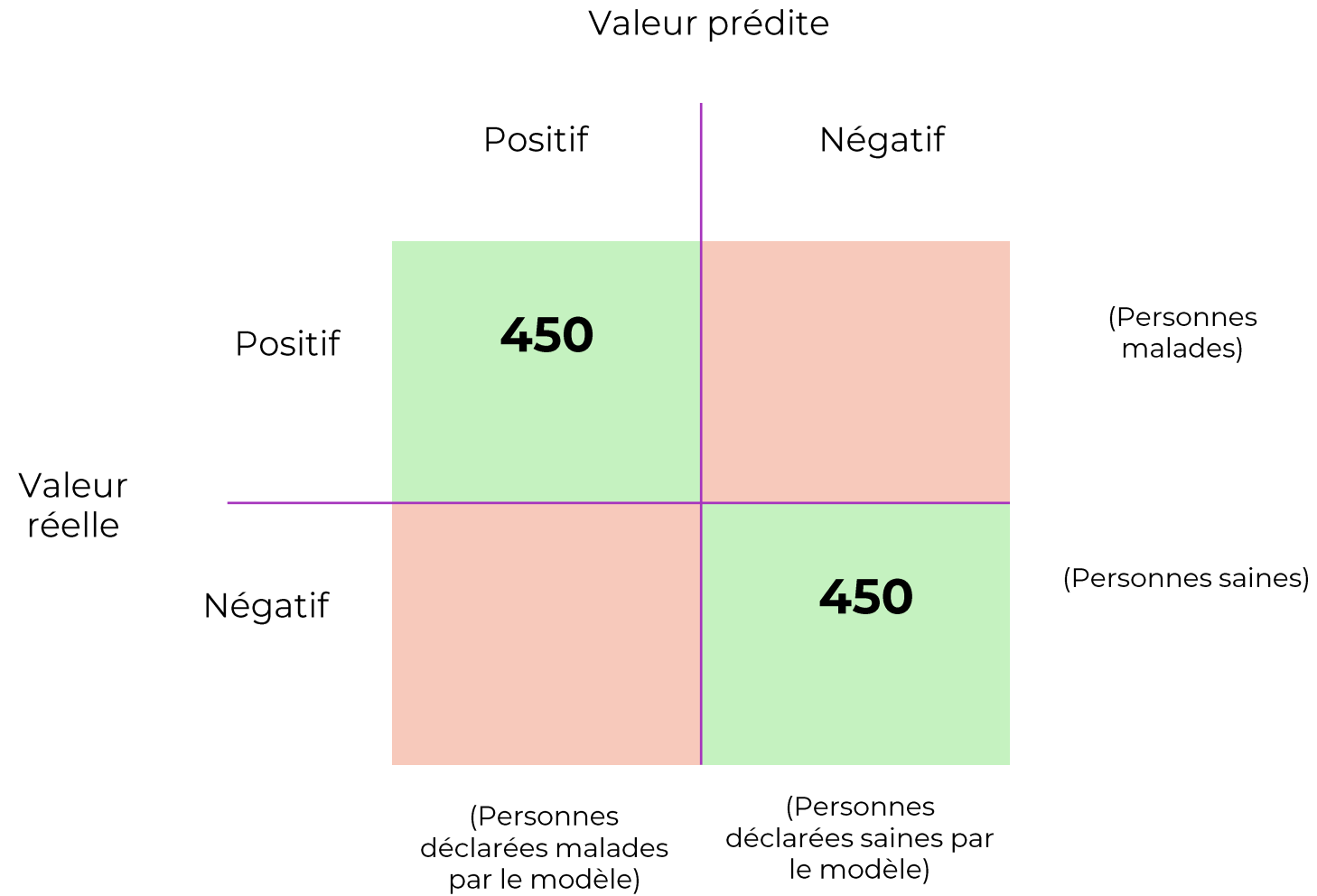


Matrice de confusion









		Valeur prédite		
		Positif	Négatif	
Valeur réelle	Positif	450	50	(Personnes malades)
	Négatif		450	(Personnes saines)
		(Personnes déclarées malades par le modèle)	(Personnes déclarées saines par le modèle)	

		Valeur prédite			
		Positif	Négatif		
Valeur réelle	Positif	450	50	(Personnes malades)	
	Négatif	50	450	(Personnes saines)	
		(Personnes déclarées malades par le modèle)	(Personnes déclarées saines par le modèle)		

		Valeur prédite	
		Positif	Négatif
Valeur réelle	Positif	450	50
	Négatif	50	450

Mesure de Performance « Accuracy »
(Acc ou Exactitude)

$$\text{Acc} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre de prédictions totales}} = \frac{900}{1000} = 0.9$$

90% du temps le modèle a raison.



6. Les modèles obtenus avec TADA



6.1 Premier modèle

MODÈLE 1

	Predicted Positive	Predicted Negative	Total
Actual Positive	48.61%	2.77%	51.38%
Actual Negative	1.38%	47.22%	48.61%
Total	50%	50%	100%

TEST METRICS

ACCURACY (ACC)



According to our test sample, this model's predictions are accurate 96% of times.

Accuracy (ACC) is a measure of the model trueness between correct predictions and all predictions. ACC interval is [0; 1]: 1 represents a perfect prediction, 0 a total inverse prediction.

VARIABLES USED

essoufflement_effort

examen_pulmo

nom_prof_sante

perte_odorat

VARIABLES USED

essoufflement_effort

examen_pulmo

nom_prof_sante

perte_odorat

Sensitivity

nom_prof_sante

62.66 %

essoufflement_effort

32.82 %

perte_odorat

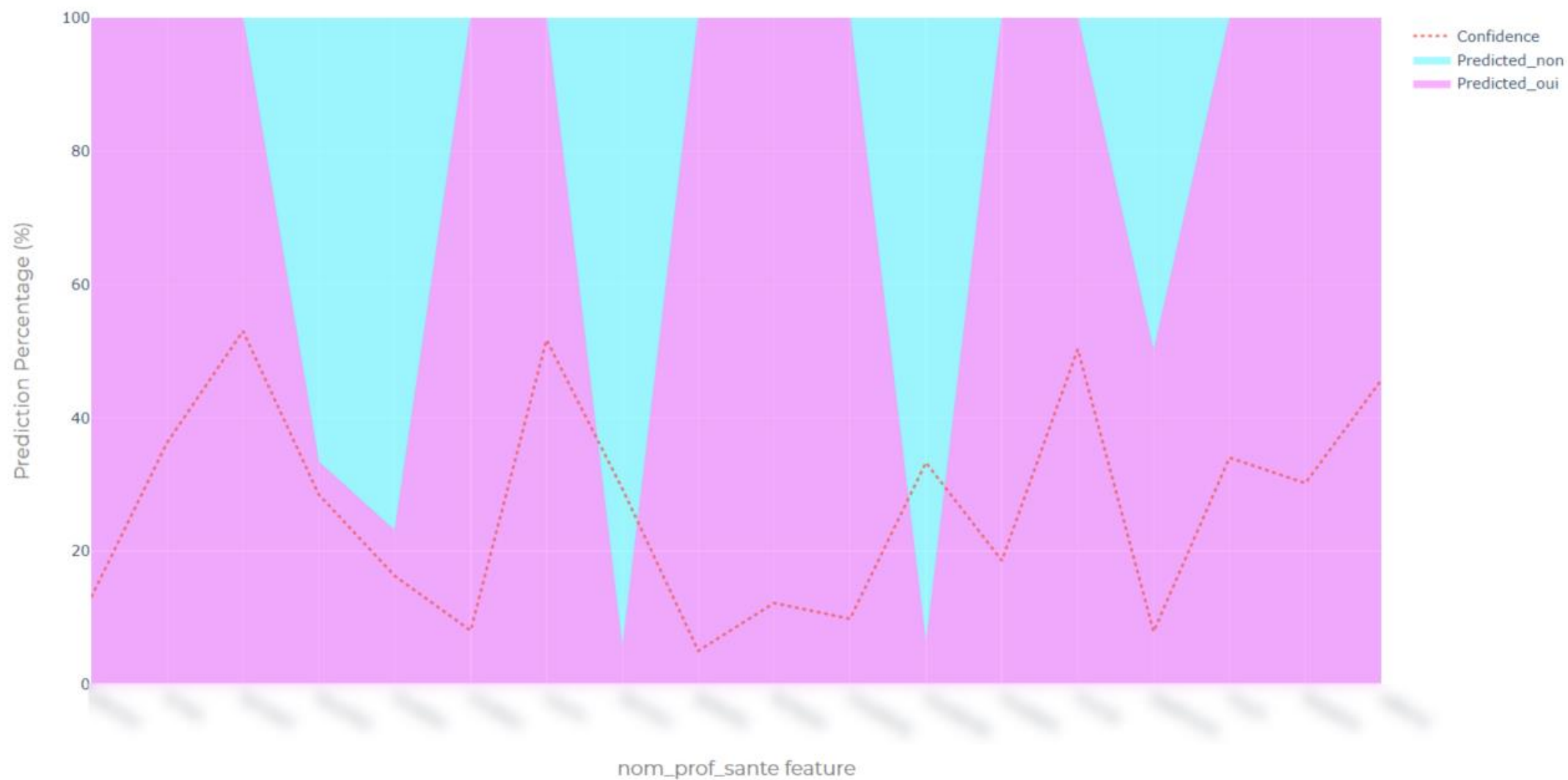
3.22 %

examen_pulmo

1.3 %

MODÈLE 1

Prediction in function of nom_prof_sante





6.2 Second modèle

	Predicted Positive	Predicted Negative	Total
Actual Positive	45.83%	2.77%	48.61%
Actual Negative	12.5%	38.88%	51.38%
Total	58.33%	41.66%	100%

TEST METRICS

ACCURACY (ACC)



According to our test sample, this model's predictions are accurate 85% of times.

Accuracy (ACC) is a measure of the model trueness between correct predictions and all predictions. ACC interval is [0; 1]: 1 represents a perfect prediction, 0 a total inverse prediction.

VARIABLES USED

desire_etre_rappele

dlr_thoraciques

essoufflement_effort

maux_de_tete

MODÈLE 2

Sensitivity

maux_de_tete

60.13 %

essoufflement_effort

21.49 %

dlr_thoraciques

14.07 %

desire_etre_rappele

4.32 %



7. Conclusion

- ➡ Nos résultats:
- Etude préliminaire très encourageante qui va être poursuivie sur l'ensemble des consultations
 - Modèle exacte 85% du temps
 - Trois symptômes clés identifiés :
Maux de tête,
essoufflement à l'effort,
douleurs thoraciques
 - 95 % des patients diagnostiqué COVID-19 par les praticiens le sont aussi par le modèle
 - Prise en charge en téléconsultation possible et isolation plus rapide.

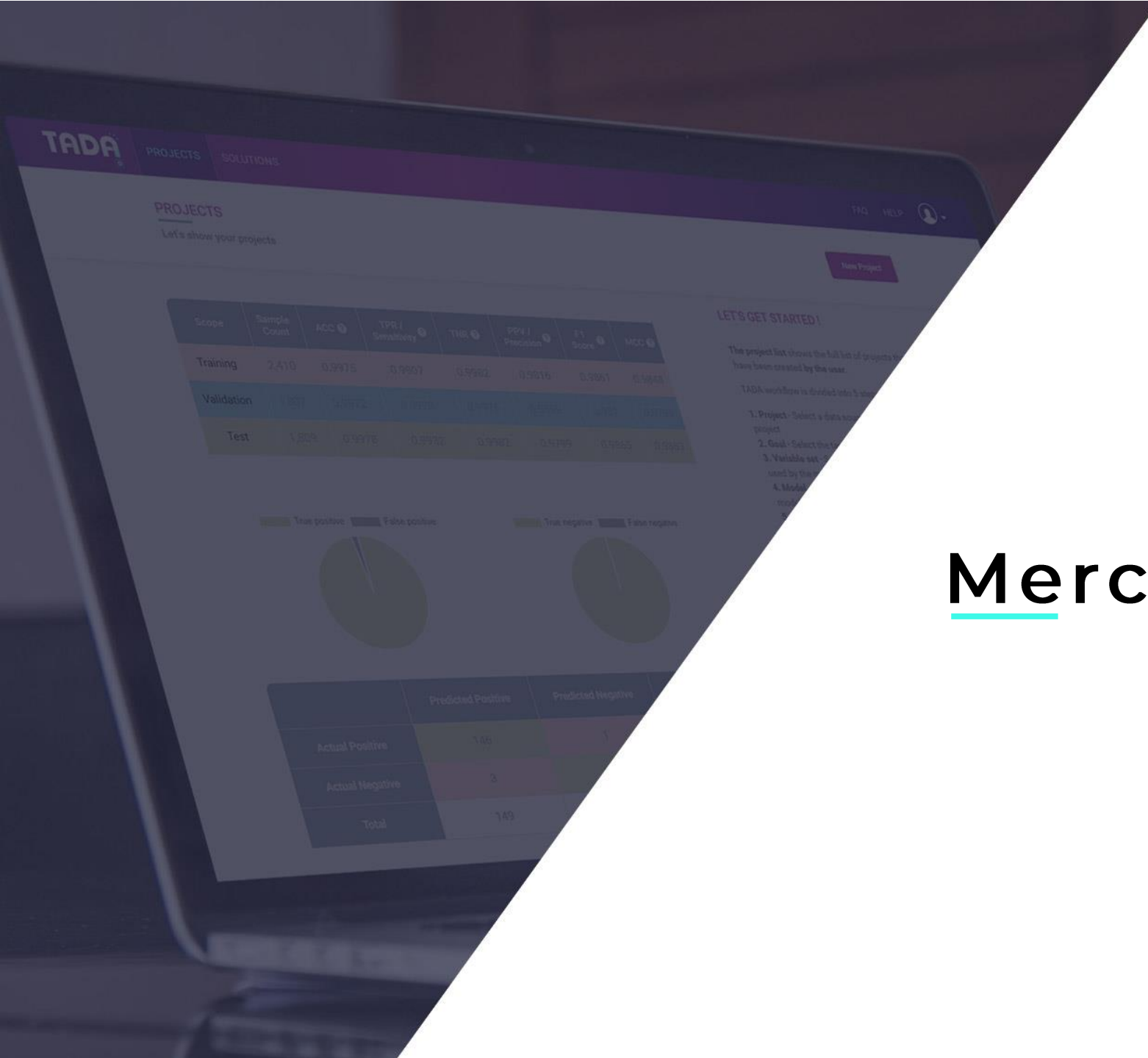
CONCLUSION

➡ Limitations:

- Jeu de données trop petit (144 observations)
- Biais d'observation des praticiens (85% de cas COVID-19)
- Besoin de valider les diagnostics des médecins avec des tests sérologiques/PCRs

➡ Perspectives:

- Dépôt d'un Projet ANR « EUREKA » (UCA/LAVAL/OMNIMED/MDM)
- Mise en place d'une architecture open-source de stockages sécurisé autonome de modèles pour faciliter le partage de nos modèles



Merci

➡ Nos résultats:

- Etude préliminaire très encourageante qui va être poursuivie sur l'ensemble des consultations
- Modèle juste 85% du temps
- Trois symptômes clés identifiés : Maux de tête, essoufflement à l'effort, douleurs thoraciques
- 95 % des patients diagnostiqué COVID-19 par les praticiens le sont aussi par le modèle
- Prise en charge en téléconsultation possible et isolation plus rapide.