
Allergen Chip Challenge

Projet T.E.R

2024-2025

Auteurs

AL AYOUBI Ibrahim
SAOUMA Christen

Encadrant

PONCELET Pascal



Nous remercions chaleureusement le Dr Michel Moise, medecin au CHU de Nîmes, pour ses explications claires lors des réunions organisées, qui nous ont permis de mieux comprendre les aspects biologiques du projet. Son expertise a grandement enrichi notre travail.

Table des matières

1	À propos du challenge	3
2	Structure et Contenu du Jeu de Données	4
2.1	Exploration des données	4
2.1.1	Informations générales et démographiques	4
2.1.2	Cibles à classifier	6
2.1.3	Features (allergènes, symptômes, etc.)	6
2.2	Ingénierie des données	7
3	Classification et Résultats	9
3.1	Rappel des Techniques et Concepts du Machine Learning	9
3.2	Classification d'allergie et sous-categories	10
3.2.1	Résultats ISAC_V1	13
3.2.2	Résultats ISAC_V2	17
3.2.3	Résultats ALEX	20
4	Analyse et Évaluation Des Modèles	23
4.1	Choix du modèle.. . . .	23
4.2	Analyse des variables les plus contributives à la classification	26
4.3	Hyperparamètres	28
5	Conclusion	29
6	Bibliographie	30

Chapitre 1

À propos du challenge

L'augmentation constante des allergies nécessite d'améliorer le diagnostic des patients, qui repose sur l'histoire clinique, des **tests cutanés** (TC) et la détection d'anticorps **IgE spécifiques** (IgEs) dans le sang. L'objectif est d'identifier ou d'exclure les allergènes potentiellement impliqués.

Nous disposons actuellement de tests IgEs multiplex basés sur des puces qui permettent une détection simultanée de centaines de spécificités allergéniques avec seulement 100 µL de sérum. Par exemple, Allergy Explorer ALEX² (Macro-Array Diagnostics, Vienne, Autriche) utilise **117 extraits allergéniques** et **178 composants protéiques**. Cet avantage à détecter les IgEs à de nombreux allergènes en un test est contrebalancé par les difficultés liées à l'interprétation.

Le data challenge ***Allergen Chip Challenge*** a été une première tentative pour résoudre cette difficulté d'interprétation. L'objectif du challenge était de développer un algorithme d'interprétation par intelligence artificielle (IA) capable de prédire la présence et la sévérité d'une maladie allergique en fonction de son profil d'IgEs obtenu par puce. Il s'agissait d'un projet collaboratif mené par la Société Française d'Allergologie (SFA) en partenariat avec le **Health Data Hub**. Ce projet financé par la banque publique d'investissement (BPI) a été mené entre décembre 2021 et novembre 2023. Pour créer ce data challenge, la SFA a coordonné la collecte des données de **4271 patients** avec leurs antécédents et les données d'IgEs avec **12 laboratoires d'allergologie** français du réseau AllergoBioNet et leurs homologues cliniciens (Bordeaux, Lille, Lyon, Marseille, Toulouse, Trousseau, Bichat, Caen, Reims, Clermont, Dijon et **Nîmes**). La base de données de l'ACC était rétrospective et incluait des données de 2012 à 2023. Les données cliniques ont été collectées dans les serveurs patient et les valeurs des IgEs dans les laboratoires effectuant les puces.

Chapitre 2

Structure et Contenu du Jeu de Données

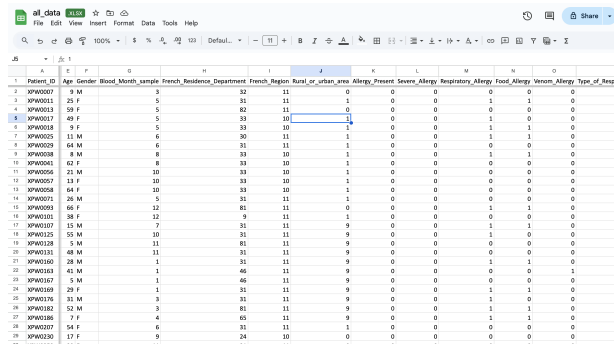
Une base de données nous a été fournie dans le cadre du projet, accompagnée d'un fichier **.xlsx** de référence.

Trois types de tests sont présents dans la base : **ISAC_V1**, **ISAC_V2** et **ALEX**. Afin de structurer notre analyse de manière claire et progressive, le projet a été divisé en trois parties, chacune correspondant à l'un de ces tests.

2.1 Exploration des données

2.1.1 Informations générales et démographiques

Dans cette première partie, nous présentons les différentes visualisations réalisées afin de mieux comprendre les données. Nous commencerons par décrire brièvement les colonnes les plus importantes, avant de passer à l'analyse visuelle des données.



Patient_ID	Age	Gender	Blood_Month_sample	French_Residence_Department	French_Region	Rural_or_urban_area	Allergy_Present	Severe_Allergy	Respiratory_Allergy	Food_Allergy	Venom_Allergy	Type_of_Respir	Type_of_Food
1	32	M	3	32	11	0	0	0	0	0	0	0	0
2	25	F	5	31	11	1	0	0	1	1	0	0	0
3	59	F	5	82	11	0	0	0	0	0	0	0	0
4	49	F	5	33	10	1	0	0	1	0	0	0	0
5	9	F	5	33	10	1	0	0	1	1	0	0	0
6	11	M	6	30	11	1	0	0	1	1	0	0	0
7	66	M	6	31	11	1	0	0	0	0	0	0	0
8	8	M	8	33	10	1	0	0	1	1	0	0	0
9	62	F	8	33	10	1	0	0	0	0	0	0	0
10	21	M	10	33	10	1	0	0	0	0	0	0	0
11	13	F	10	33	10	1	0	0	0	0	0	0	0
12	26	M	5	31	11	1	0	0	0	0	0	0	0
13	64	F	10	33	10	1	0	0	0	0	0	0	0
14	66	F	12	81	11	0	0	0	1	1	0	0	0
15	38	F	12	9	11	1	0	0	0	0	0	0	0
16	15	M	7	31	11	9	0	0	1	1	0	0	0
17	55	M	10	31	11	9	0	0	1	0	0	0	0
18	5	M	11	81	11	9	0	0	0	0	0	0	0
19	48	M	11	31	11	9	0	0	0	0	0	0	0
20	28	M	1	31	11	9	0	0	1	1	0	0	0
21	45	M	1	46	11	9	0	0	0	0	0	0	0
22	5	M	1	46	11	9	0	0	0	0	0	0	0
23	29	F	1	31	11	9	0	0	1	1	0	0	0
24	31	M	3	31	11	9	0	0	1	0	0	0	0
25	52	M	3	81	11	9	0	0	1	1	0	0	0
26	7	F	4	65	11	9	0	0	1	1	0	0	0
27	34	F	6	31	11	9	0	0	0	0	0	0	0
28	17	F	9	24	10	0	0	0	0	0	0	0	0

FIGURE 2.1 – Jeu de Données.

Chaque ligne du fichier correspond à une observation unique, identifiée par un `Patient_ID`, et associée à un `Chip_ID` représentant le type de test effectué.

Outre l'identifiant du patient et le type de test, plusieurs informations contextuelles sont fournies, telles que :

- le **genre** du patient (masculin ou féminin),
- l'**âge** au moment du test,
- le **mois de la prise de sang**,
- le **département** de résidence,
- la **region** française,
- le **type d'habitat** (urbain ou rural, par exemple).

Ces variables offrent un aperçu global du profil des patients et permettent de mieux comprendre la distribution des tests dans différents contextes géographiques et temporels. Une première visualisation a été réalisée à partir de ces attributs, présentée ci-dessous afin d'illustrer cette diversité.

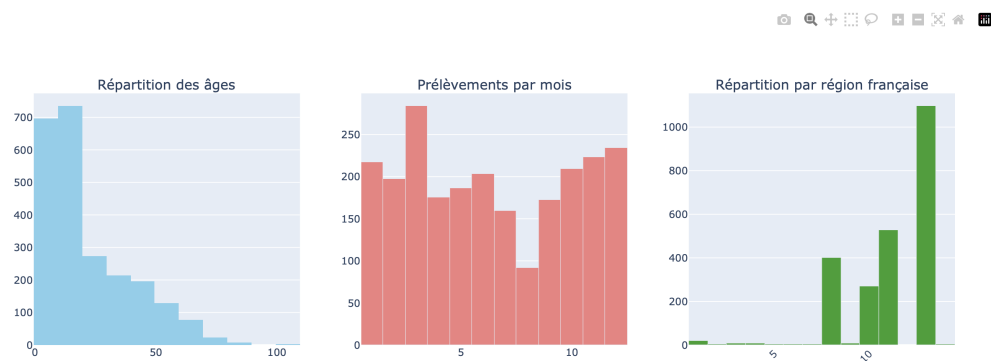


FIGURE 2.2 – Distribution des patients selon l'âge, le département et les prélèvement par mois.

Comme on peut le voir dans les graphiques, on constate à partir de ces premières visualisations que la majorité des personnes ayant effectué les tests ont moins de **20 ans**. Le mois où le plus de tests ont été réalisés est mars. De même, dans la **région 13**, on observe un taux très élevé de personnes ayant effectué le test V1.

2.1.2 Cibles à classifier

Dans une seconde phase de l'analyse, les autres colonnes du dataset correspondent aux **variables cibles** (targets) ainsi qu'à des caractéristiques plus spécifiques. Les colonnes associées aux cibles sont binaires (valeurs 0 ou 1), indiquant la présence ou l'absence d'un certain type d'allergie chez le patient. Les variables cibles principales incluent : *Allergy_Present*, *Respiratory_Allergy*, *Food_Allergy*, *Venom_Allergy* et *Severe_Allergy*.

Ces catégories globales sont ensuite détaillées à travers des sous-types, permettant une granularité plus fine de l'analyse. Par exemple, dans la catégorie des allergies respiratoires, on retrouve des indicateurs spécifiques pour les allergies au pollen, à l'ARIA et à la poussière (Conv). De même, les allergies alimentaires sont subdivisées en plusieurs entités, telles que les allergies aux œufs, au poisson, aux légumes, etc. Cette structuration hiérarchique des variables permet de réaliser des classifications plus détaillées des profils allergiques.

2.1.3 Features (allergènes, symptômes, etc.)

Profil allergénique

Le test **ISAC_V1** permet d'identifier une large gamme d'allergènes répartis en différentes classes biologiques. Les allergènes d'origine *pollinique* incluent notamment *Bet_v_1*, *Phl_p_1*, *Cup_a_1* ou *Pla_a_3*. Les *moisissures* comme *Alt_a_1* ou *Asp_f_1* sont également représentées. Les allergènes d'origine *animale* comprennent les épithéliums de chat (*Fel_d_1*), de chien (*Can_f_1*), ou encore les acariens tels que *Der_p_1* et *Blo_t_5*. On retrouve également des allergènes *alimentaires* très variés, comme ceux de l'arachide (*Ara_h_1* à *Ara_h_9*), des fruits à coque (*Jug_r_1*, *Cor_a_9*), des produits laitiers (*Bos_d_8*), ou encore des poissons (*Gad_c_1*). Les *venins d'insectes* (*Api_m_1*, *Ves_v_5*) sont également couverts. Enfin, certaines structures *glycosylées* comme *MUXF3* ou *Hev_b_8* sont identifiées, représentant des motifs pan-allergéniques.

Cependant, le test **ISAC_V1** présente certaines limitations, notamment un nombre d'allergènes restreint par rapport à la version suivante **ISAC_V2**. Cette dernière a permis l'ajout de plusieurs allergènes pertinents, comme *Can_f_4*, *Can_f_6*, *Der_p_23*, ou encore *Ana_o_3*, améliorant ainsi la couverture diagnostique, notamment pour certaines allergies spécifiques aux animaux ou aux insectes. Néanmoins, malgré ces ajouts, la capacité d'ISAC_V2 reste limitée en termes de nombre total d'allergènes et de diversité des familles représentées. Le test **ALEX** (Allergy Explorer) constitue une avancée majeure dans le domaine du diagnostic allergologique. En effet, il propose une **couverture plus large et plus complète** que les

tests ISAC_V1 et ISAC_V2, incluant à la fois des allergènes moléculaires et des extraits complets. ALEX permet de tester simultanément plus de 280 allergènes, couvrant ainsi une gamme beaucoup plus étendue, incluant des allergènes environnementaux, alimentaires, et des venins. De plus, ALEX intègre des allergènes non présents dans ISAC, comme des allergènes de latex spécifiques (Hev_b_6.02), des composants de fruits de mer, ou encore des extraits de protéines animales spécifiques (par exemple Bos_d_meat, Bos_d_milk, Equ_c_milk).

Symptômes, facteurs cliniques et traitements

Passons maintenant aux variables explicatives (**features**) présentes dans ce dataset. Celui-ci contient plusieurs colonnes décrivant différents types de symptômes, tels que les symptômes cardiovasculaires, les symptômes cutanés, ainsi que des informations relatives aux traitements contre la rhinite ou l'asthme, et d'autres facteurs généraux. Le dataset comporte un nombre important de variables, mais également de **nombreuses valeurs manquantes**. C'est pourquoi, dans une section suivante, nous détaillerons les différentes étapes d'ingénierie des données que nous avons mises en œuvre afin d'obtenir des features mieux structurées et adaptées aux tâches de classification.

2.2 Ingénierie des données

Cette partie a représenté le défi le plus complexe du projet, en raison du grand nombre de tâches à effectuer. Tout d'abord, nous avons séparé les colonnes relatives aux traitements de la rhinite et de l'asthme. Ces colonnes contenaient des valeurs allant de 0 à 5 (les significations de ces valeurs sont précisées dans le dictionnaire). Pour chaque valeur possible, j'ai créé une nouvelle colonne ; si la valeur correspondante était présente, j'y attribuais 1, sinon 0. Le même type de traitement a été appliqué aux colonnes *general_cofactor*, *Age_of_onsets*, etc.

La figure ci-dessous offre un aperçu des nouvelles colonnes créées.

Chapitre 3

Classification et Résultats

3.1 Rappel des Techniques et Concepts du Machine Learning

Voici un bref rappel des modèles utilisés :

1. **Random Forest** : Il s'agit d'un modèle d'ensemble basé sur les arbres de décision. Il construit plusieurs arbres sur des sous-échantillons des données d'entraînement et combine leurs prédictions (par vote majoritaire pour la classification) pour améliorer la performance globale et réduire le surapprentissage.
2. **XGBoost (Extreme Gradient Boosting)** : C'est un algorithme de boosting qui construit les arbres de manière séquentielle, chaque nouvel arbre cherchant à corriger les erreurs des précédents. Il repose sur le principe du gradient boosting, en optimisant une fonction de perte à l'aide de la descente de gradient.
3. **SVM (Support Vector Machine)** : C'est un modèle linéaire (ou non linéaire via le noyau) qui cherche à trouver l'hyperplan optimal séparant les classes avec la plus grande marge. Il est particulièrement efficace pour les problèmes de classification binaire.
4. **Régression Logistique** : Modèle linéaire utilisé pour la classification. Il estime la probabilité qu'un échantillon appartienne à une classe en utilisant la fonction sigmoïde. Il est simple, interprétable, et souvent utilisé comme modèle de base.

Quelques notions utilisés :

1. **K-Fold Cross-Validation** : est une méthode de validation croisée qui consiste à diviser l'ensemble de données en K sous-ensembles égaux. À chaque itération, un sous-ensemble est utilisé comme ensemble de test, tandis que les

autres servent à l'entraînement. Cette technique permet d'évaluer la performance d'un modèle de manière plus robuste et d'éviter le surapprentissage.

2. **AUC - ROC (Area Under the Curve - Receiver Operating Characteristic)** : est une métrique d'évaluation qui mesure la capacité d'un modèle à distinguer entre classes positives et négatives. La courbe ROC trace le Taux de vrais positifs (TPR) contre le Taux de faux positifs (FPR). Plus l'AUC est proche de 1, meilleur est le modèle en termes de discrimination.
3. **Hyperparamètres** : sont des paramètres définis avant l'entraînement d'un modèle et influencent directement son comportement. Ils contrôlent des aspects tels que la profondeur d'un arbre de décision, le taux d'apprentissage ou le nombre de voisins dans un KNN. Trouver les bons hyperparamètres est crucial pour obtenir un modèle performant.

3.2 Classification d'allergie et sous-categories

Nous allons présenter les classifications effectuées pour chaque test et chaque cible. Tout d'abord, notre méthode consistait à classer les patients en deux catégories : patients allergiques et non allergiques. Ensuite, parmi les patients allergiques, nous avons cherché à déterminer s'il était possible de les classer selon le type d'allergie : respiratoire ou non, alimentaire ou non, venin ou non. Dans une autre étape, nous avons affiné cette classification en distinguant les sous-catégories, comme respiratoire et alimentaire, en utilisant la même approche. Enfin, dans la dernière partie, nous avons évalué la sévérité de l'allergie en effectuant une classification spécifique sur ce critère.

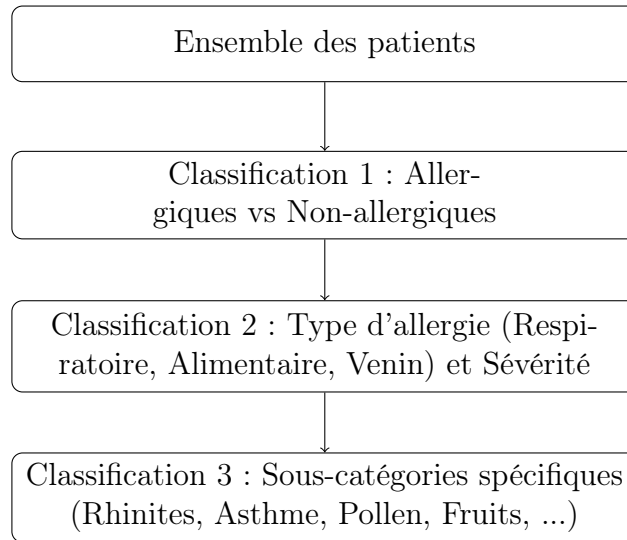


FIGURE 3.1 – Étapes de classification des allergies et sous-catégories

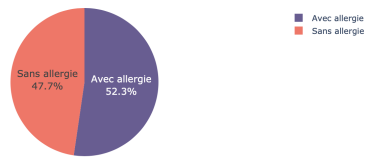
Avant d’entraîner les modèles de classification, une première étape essentielle a consisté à vérifier l’équilibrage des données. Un jeu de données déséquilibré peut fortement biaiser les résultats des algorithmes, les rendant moins performants pour prédire les classes minoritaires. Pour cela, la distribution des différentes classes (présence ou absence d’allergie) a été analysée afin de s’assurer que le modèle ne favorise pas une classe au détriment des autres.

Dans notre approche, nous avons considéré deux cas distincts :

- Le premier cas sans équilibrage, où les données sont utilisées telles quelles, afin d’observer l’impact du déséquilibre sur les performances des modèles.
- Le second cas avec équilibrage des classes, nous avons appliqué la méthode SMOTE (Synthetic Minority Over-sampling Technique), qui permet de générer artificiellement des exemples de la classe minoritaire.

Ci-dessous, nous présentons des graphiques illustrant la distribution des classes pour différents tests et différentes cibles, afin de mieux comprendre l’état des données avant l’entraînement des modèles.

Répartition des classes dans 'Allergy_Present'



Allergie vs Non Allergie (V1)

Répartition des classes dans 'Allergy_Present'



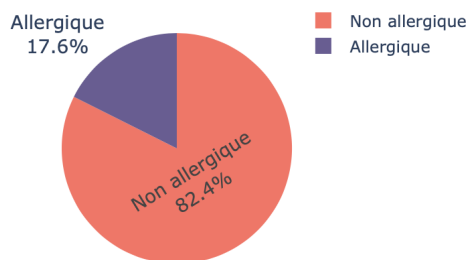
Allergie vs Non Allergie (V2)

Répartition des classes dans 'Allergy_Present'



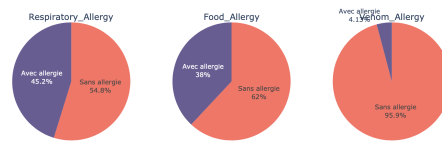
Allergie vs Non Allergie (ALEX)

Type_of_Food_Allergy_Fruits_and_Vegetables



Allergie aux fruits et legumes (ALEX)

Répartition des allergies par type



Respiratoire - Alimentaire - Venin (V1)

Répartition des allergies par type



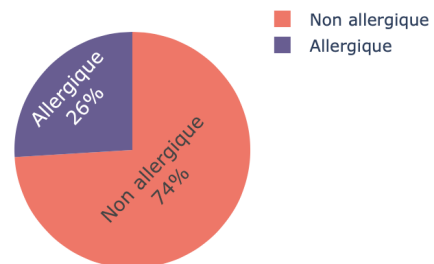
Respiratoire - Alimentaire - Venin (V2)

Répartition des allergies par type



Respiratoire - Alimentaire- Venin (ALEX)

Type_of_Respiratory_Allergy_CONJ



Allergie Conj (V1)

Après cette analyse préliminaire des classes et l'application éventuelle de la méthode d'équilibrage SMOTE, nous avons procédé à l'entraînement des modèles en utilisant une validation croisée **K-Fold** avec **k=10**.

Cette méthode permet d'évaluer de manière robuste les performances des modèles en s'assurant que chaque observation du jeu de données est utilisée à la fois pour l'entraînement et la validation, garantissant ainsi une meilleure généralisation des résultats.

Dans les section suivantes, nous présenterons en détail les performances obtenues pour chaque cible et chaque modèle, en nous basant sur les différents critères d'évaluation (F1-score, précision, matrices de confusion, courbes AUC-ROC, etc.).

3.2.1 Résultats ISAC_V1

Target	Model	F1-Class 1	Precision	AUC-ROC
Allergy_Present	RandomForest	0.9516	0.9501	0.9832
	XGBoost	0.9538	0.9532	0.9842
	LogisticRegression	0.9515	0.9524	0.9798
	SVM	0.8476	0.8683	0.9425
Food_Allergy	RandomForest	0.8888	0.9169	0.9690
	XGBoost	0.8829	0.9134	0.9636
	LogisticRegression	0.8168	0.8661	0.9298
	SVM	0.7315	0.8103	0.8867
Respiratory_Allergy	RandomForest	0.9603	0.9645	0.9895
	XGBoost	0.9647	0.9689	0.9909
	LogisticRegression	0.9124	0.9247	0.9760
	SVM	0.8144	0.8422	0.9150
Venom_Allergy	RandomForest	0.7691	0.9835	0.9423
	XGBoost	0.8157	0.9862	0.9539
	LogisticRegression	0.3495	0.9572	0.8684
	SVM	0.0000	0.9192	0.7719
Severe_Allergy	RandomForest	0.8771	0.8528	0.9318
	XGBoost	0.8767	0.8484	0.9258
	LogisticRegression	0.8527	0.8359	0.9183
	SVM	0.7102	0.6889	0.7213

TABLE 3.1 – Performance des modèles pour la classification des allergies (V1).

Target	Model	F1-Class 1	Precision	AUC-ROC
ATCD_Venom	RandomForest	0.6500	0.9175	0.9125
	XGBoost	0.6200	0.9061	0.9062
	LogisticRegression	0.4867	0.8507	0.8639
	SVM	0.4655	0.8720	0.9028
IGE_Venom	RandomForest	0.9947	0.9810	1.0000
	XGBoost	1.0000	1.0000	1.0000
	LogisticRegression	0.9198	0.8786	0.9042
	SVM	0.7755	0.8740	0.8458

TABLE 3.2 – Performance des modèles pour les sous types d’allergies au venin.

Target	Model	F1-Class 1	Precision	AUC-ROC
Pollen_Herb	RandomForest	0.7081	0.7456	0.8438
	XGBoost	0.7289	0.7556	0.8461
	LogisticRegression	0.7098	0.7522	0.8320
	SVM	0.6800	0.7166	0.7858
Pollen_Tree	RandomForest	0.9317	0.8980	0.9353
	XGBoost	0.9321	0.8972	0.9321
	LogisticRegression	0.8745	0.8176	0.8584
	SVM	0.8310	0.8337	0.8905
Dander_Animals	RandomForest	0.8531	0.8268	0.8852
	XGBoost	0.8379	0.8086	0.8650
	LogisticRegression	0.7363	0.7282	0.7953
	SVM	0.7523	0.7463	0.8254
Mite_Cockroach	RandomForest	0.8656	0.8465	0.9246
	XGBoost	0.8786	0.8644	0.9362
	LogisticRegression	0.7988	0.8222	0.8859
	SVM	0.7633	0.7964	0.8822
Molds_Yeast	RandomForest	0.8441	0.9092	0.9610
	XGBoost	0.8490	0.9113	0.9687
	LogisticRegression	0.6443	0.7962	0.8110
	SVM	0.7037	0.8259	0.8923
Allergy_ARIA	RandomForest	0.9921	0.9926	0.9984
	XGBoost	0.9921	0.9926	0.9990
	LogisticRegression	0.9891	0.9898	0.9996
	SVM	0.6649	0.6707	0.7263
CONJ	RandomForest	0.9836	0.9916	0.9997
	XGBoost	0.9982	0.9991	0.9997
	LogisticRegression	0.6295	0.8060	0.8480
	SVM	0.5949	0.8104	0.7328
Pollen_Gram	RandomForest	0.8672	0.8277	0.9112
	XGBoost	0.8835	0.8530	0.9335
	LogisticRegression	0.8292	0.8063	0.8787
	SVM	0.7974	0.7852	0.8563
GINA	RandomForest	0.9907	0.9907	0.9991
	XGBoost	0.9887	0.9890	0.9994
	LogisticRegression	0.9878	0.9881	0.9996
	SVM	0.6614	0.6733	0.7326

TABLE 3.3 – Performance des modèles pour les sous types respiratoire.

Target	Model	F1-Class 1	Precision	AUC-ROC
Aromatics	RandomForest	0.0667	0.9662	0.8009
	XGBoost	0.1900	0.9727	0.7527
	LogisticRegression	0.0832	0.9684	0.7121
	SVM	0.0868	0.9737	0.7221
Cereals_&_Seeds	RandomForest	0.0000	0.9403	0.7198
	XGBoost	0.0000	0.9400	0.6871
	LogisticRegression	0.0832	0.9447	0.5341
	SVM	0.1009	0.9522	0.5536
Egg	RandomForest	0.0667	0.9707	0.9023
	XGBoost	0.3500	0.9800	0.9335
	LogisticRegression	0.0933	0.9726	0.7773
	SVM	0.0520	0.9716	0.6449
Fish	RandomForest	0.0000	0.9557	0.6779
	XGBoost	0.0667	0.9589	0.6399
	LogisticRegression	0.0667	0.9589	0.5502
	SVM	0.0838	0.9626	0.6704
Fruits_and_Vegetables	RandomForest	0.0733	0.9278	0.8438
	XGBoost	0.2305	0.9409	0.8613
	LogisticRegression	0.2003	0.9395	0.7425
	SVM	0.2224	0.9499	0.8021
Oral_Syndrom	RandomForest	0.8815	0.9742	0.9989
	XGBoost	0.9960	0.9990	0.9995
	LogisticRegression	0.4118	0.8469	0.8056
	SVM	0.3789	0.8590	0.7192
Other_Legumes	RandomForest	0.0000	0.9646	0.7300
	XGBoost	0.0667	0.9662	0.6642
	LogisticRegression	0.0686	0.9667	0.5197
	SVM	0.0473	0.9677	0.5754
Peanut	RandomForest	0.2453	0.8940	0.8190
	XGBoost	0.3065	0.8972	0.8573
	LogisticRegression	0.2537	0.8876	0.6785
	SVM	0.2782	0.9020	0.7270
Shellfish	RandomForest	0.0500	0.9450	0.7831
	XGBoost	0.2000	0.9547	0.8008
	LogisticRegression	0.1430	0.9500	0.7111
	SVM	0.1352	0.9566	0.7325
TPO	RandomForest	0.2317	0.9247	0.8501
	XGBoost	0.2683	0.9277	0.8227
	LogisticRegression	0.2230	0.9231	0.7880
	SVM	0.2910	0.9433	0.8249
Tree_Nuts	RandomForest	0.2820	0.8992	0.8458
	XGBoost 16	0.4085	0.9070	0.8562
	LogisticRegression	0.2944	0.8830	0.7658
	SVM	0.2641	0.8922	0.7194

TABLE 3.4 – Performance des modèles pour les sous types d’allergies alimentaires.

3.2.2 Résultats ISAC_V2

Target	Model	F1-Class 1	Precision	AUC-ROC
Allergy_Present	RandomForest	0.9923	0.9900	0.9986
	XGBoost	0.9904	0.9876	0.9997
	LogisticRegression	0.9807	0.9756	0.9956
	SVM	0.9626	0.9500	0.9650
Respiratory_Allergy	RandomForest	0.9887	0.9864	0.9993
	XGBoost	0.9886	0.9862	0.9980
	LogisticRegression	0.9536	0.9435	0.9896
	SVM	0.9272	0.9132	0.9480
Food_Allergy	RandomForest	0.9482	0.9577	0.9904
	XGBoost	0.9381	0.9483	0.9859
	LogisticRegression	0.8886	0.9069	0.9525
	SVM	0.8206	0.8612	0.9369
Severe_Allergy	RandomForest	0.9707	0.9558	0.9838
	XGBoost	0.9643	0.9460	0.9835
	LogisticRegression	0.9390	0.9109	0.9571
	SVM	0.7064	0.8003	0.7994

TABLE 3.5 – Performance des modèles pour la classification des allergies (V2).

Target	Model	F1-Class 1	Precision	AUC-ROC
Pollen_Herb	RandomForest	0.6639	0.7902	0.8671
	XGBoost	0.6903	0.8052	0.8659
	LogisticRegression	0.6594	0.8015	0.8256
	SVM	0.6305	0.7645	0.8063
Pollen_Tree	RandomForest	0.9139	0.9121	0.9640
	XGBoost	0.8937	0.8855	0.9508
	LogisticRegression	0.8507	0.8508	0.9234
	SVM	0.8210	0.8339	0.9234
Dander_Animals	RandomForest	0.7934	0.8118	0.9040
	XGBoost	0.7955	0.8094	0.8788
	LogisticRegression	0.7158	0.7476	0.8355
	SVM	0.7702	0.7799	0.8453
Mite_Cockroach	RandomForest	0.8996	0.8802	0.9282
	XGBoost	0.9021	0.8848	0.9378
	LogisticRegression	0.8156	0.8018	0.8872
	SVM	0.7921	0.8018	0.9018
Molds_Yeast	RandomForest	0.8212	0.9126	0.9439
	XGBoost	0.8292	0.9155	0.9537
	LogisticRegression	0.4754	0.7497	0.7446
	SVM	0.5760	0.7857	0.8482
Allergy_ARIA	RandomForest	0.9906	0.9881	0.9999
	XGBoost	0.9952	0.9941	0.9993
	LogisticRegression	0.9937	0.9920	1.0000
	SVM	0.5995	0.6678	0.6896
Allergy_CONJ	RandomForest	0.9773	0.9869	0.9996
	XGBoost	0.9968	0.9980	1.0000
	LogisticRegression	0.8733	0.9254	0.9456
	SVM	0.4178	0.6517	0.6653
Pollen_Gram	RandomForest	0.8548	0.8408	0.9220
	XGBoost	0.8733	0.8629	0.9473
	LogisticRegression	0.7796	0.7923	0.8850
	SVM	0.7935	0.7903	0.8703
Allergy_GINA	RandomForest	0.9877	0.9843	0.9958
	XGBoost	0.9859	0.9821	0.9966
	LogisticRegression	0.9858	0.9820	0.9985
	SVM	0.7677	0.6685	0.7011

TABLE 3.6 – Performance des modèles pour les sous-types respiratoires (V2).

Target	Model	F1-Class 1	Precision	AUC-ROC
Aromatics	RandomForest	0.1067	0.8693	0.5674
	XGBoost	0.0250	0.8552	0.5498
	LogisticRegression	0.0222	0.8522	0.6520
	SVM	0.1579	0.8801	0.6454
Cereals_&_Seeds	RandomForest	0.0000	0.8421	0.4171
	XGBoost	0.0000	0.8399	0.4043
	LogisticRegression	0.0619	0.8476	0.5114
	SVM	0.1448	0.8548	0.5124
Egg	RandomForest	0.1472	0.8002	0.6057
	XGBoost	0.2376	0.8205	0.6647
	LogisticRegression	0.2275	0.8074	0.6290
	SVM	0.2121	0.7991	0.5370
Fish	RandomForest	0.2605	0.8550	0.5580
	XGBoost	0.2110	0.8327	0.5347
	LogisticRegression	0.1321	0.8098	0.6150
	SVM	0.2238	0.8350	0.5585
Fruits_and_Vegetables	RandomForest	0.3840	0.7107	0.6718
	XGBoost	0.4243	0.7211	0.7118
	LogisticRegression	0.3690	0.6861	0.6629
	SVM	0.3587	0.6521	0.5927
Mammalian_Milk	RandomForest	0.0000	0.8932	0.7111
	XGBoost	0.0000	0.8916	0.7056
	LogisticRegression	0.0619	0.8981	0.6160
	SVM	0.2175	0.9325	0.7441
Other_Legumes	RandomForest	0.2137	0.7785	0.6225
	XGBoost	0.2012	0.7737	0.6024
	LogisticRegression	0.2833	0.7895	0.6460
	SVM	0.2366	0.7539	0.5172
Peanut	RandomForest	0.4479	0.6897	0.6749
	XGBoost	0.4233	0.6727	0.6585
	LogisticRegression	0.3807	0.6401	0.6567
	SVM	0.3577	0.6417	0.5924
Shellfish	RandomForest	0.0000	0.8071	0.7018
	XGBoost	0.2160	0.8463	0.7407
	LogisticRegression	0.1639	0.8343	0.5811
	SVM	0.2118	0.8384	0.6328
Tree_Nuts	RandomForest	0.5770	0.6799	0.7386
	XGBoost	0.5610	0.6659	0.7208
	LogisticRegression	0.5671	0.6684	0.6925
	SVM	0.4406	0.5684	0.5938

TABLE 3.7 – Performance des modèles pour les sous-types d’allergies alimentaires (V2).

3.2.3 Résultats ALEX

Target	Model	F1-Class 1	Precision	AUC-ROC
Allergy_Present	RandomForest	0.9718	0.9717	0.9955
	XGBoost	0.9614	0.9611	0.9901
	LogisticRegression	0.9467	0.9485	0.9814
	SVM	0.8804	0.8753	0.9319
Respiratory_Allergy	RandomForest	0.9451	0.9530	0.9891
	XGBoost	0.9470	0.9551	0.9868
	LogisticRegression	0.9158	0.9287	0.9705
	SVM	0.8062	0.8323	0.9098
Food_Allergy	RandomForest	0.8641	0.9087	0.9691
	XGBoost	0.8739	0.9162	0.9714
	LogisticRegression	0.7762	0.8487	0.9080
	SVM	0.7659	0.8436	0.9048
Venom_Allergy	RandomForest	0.3513	0.9023	0.8840
	XGBoost	0.3829	0.9049	0.8575
	LogisticRegression	0.2575	0.8837	0.6994
	SVM	0.2312	0.8860	0.7180
Severe_Allergy	RandomForest	0.8936	0.8368	0.8860
	XGBoost	0.8804	0.8217	0.8845
	LogisticRegression	0.8667	0.8319	0.8610
	SVM	0.7232	0.7136	0.6494

TABLE 3.8 – Performance des modèles pour la classification des allergies (ALEX).

Target	Model	F1-Class 1	Precision	AUC-ROC
ATCD_Venom	RandomForest	0.0667	0.7538	0.6384
	XGBoost	0.2000	0.7800	0.7232
	LogisticRegression	0.2667	0.7885	0.6125
	SVM	0.1900	0.7529	0.5357

TABLE 3.9 – Performance des modèles pour la classification des allergies au venin (ALEX).

Target	Model	F1-Class 1	Precision	AUC-ROC
Pollen_Herb	RandomForest	0.7520	0.7565	0.8329
	XGBoost	0.7583	0.7619	0.8232
	LogisticRegression	0.6838	0.7043	0.7811
	SVM	0.6856	0.7028	0.8029
Pollen_Tree	RandomForest	0.8951	0.8725	0.9462
	XGBoost	0.9083	0.8867	0.9509
	LogisticRegression	0.8431	0.8222	0.8702
	SVM	0.8027	0.8022	0.8862
Dander_Animals	RandomForest	0.9064	0.9060	0.9491
	XGBoost	0.9088	0.9044	0.9428
	LogisticRegression	0.8268	0.8351	0.8824
	SVM	0.7808	0.8038	0.8714
Mite_Cockroach	RandomForest	0.9715	0.9694	0.9908
	XGBoost	0.9786	0.9771	0.9879
	LogisticRegression	0.9422	0.9393	0.9811
	SVM	0.8237	0.8668	0.9418
Molds_Yeast	RandomForest	0.8734	0.9258	0.9428
	XGBoost	0.9020	0.9424	0.9499
	LogisticRegression	0.7301	0.8470	0.8622
	SVM	0.6870	0.8403	0.8768
ARIA	RandomForest	0.9790	0.9757	0.9983
	XGBoost	1.0000	1.0000	1.0000
	LogisticRegression	0.9933	0.9928	1.0000
	SVM	0.7016	0.6473	0.6669
CONJ	RandomForest	0.9122	0.9541	0.9940
	XGBoost	1.0000	1.0000	1.0000
	LogisticRegression	0.5221	0.7273	0.7483
	SVM	0.5080	0.7085	0.6739
Pollen_Gram	RandomForest	0.9558	0.9468	0.9746
	XGBoost	0.9425	0.9308	0.9778
	LogisticRegression	0.8730	0.8736	0.9188
	SVM	0.8336	0.8564	0.9314
GINA	RandomForest	0.8939	0.9204	0.9726
	XGBoost	0.9062	0.9289	0.9717
	LogisticRegression	0.8983	0.9234	0.9802
	SVM	0.5313	0.6309	0.6562

TABLE 3.10 – Performance des modèles pour les sous-types respiratoires (ALEX).

Target	Model	F1-Class 1	Precision	AUC-ROC
Aromatics	RandomForest	0.0400	0.8756	0.7442
	XGBoost	0.1733	0.8936	0.7593
	LogisticRegression	0.1436	0.8857	0.7171
	SVM	0.1400	0.8872	0.6352
Egg	RandomForest	0.1500	0.8756	0.6691
	XGBoost	0.1550	0.8712	0.6938
	LogisticRegression	0.2260	0.8796	0.6275
	SVM	0.1990	0.8789	0.5954
Fish	RandomForest	0.3871	0.9154	0.7352
	XGBoost	0.4152	0.9162	0.7410
	LogisticRegression	0.2415	0.8818	0.6347
	SVM	0.1530	0.8626	0.5848
Fruits_and_Vegetables	RandomForest	0.2698	0.7634	0.6587
	XGBoost	0.2745	0.7627	0.6096
	LogisticRegression	0.2570	0.7404	0.6282
	SVM	0.3523	0.7689	0.6283
Mammalian_Milk	RandomForest	0.2333	0.9210	0.7625
	XGBoost	0.2805	0.9233	0.7304
	LogisticRegression	0.2319	0.9165	0.6832
	SVM	0.1208	0.8983	0.5213
Oral_Syndrom	RandomForest	0.7588	0.9655	0.9906
	XGBoost	1.0000	1.0000	1.0000
	LogisticRegression	0.2370	0.8508	0.6591
	SVM	0.1881	0.8421	0.5967
Other_Legumes	RandomForest	0.0000	0.8490	0.6398
	XGBoost	0.0667	0.8574	0.5534
	LogisticRegression	0.0708	0.8532	0.4315
	SVM	0.1298	0.8570	0.4792
Peanut	RandomForest	0.1999	0.7751	0.6600
	XGBoost	0.1577	0.7622	0.6783
	LogisticRegression	0.2661	0.7821	0.6446
	SVM	0.2527	0.7750	0.5908
Shellfish	RandomForest	0.0500	0.8746	0.5844
	XGBoost	0.2333	0.9015	0.6360
	LogisticRegression	0.0867	0.8714	0.5185
	SVM	0.2070	0.8922	0.6400
TPO	RandomForest	0.0000	0.8891	0.4627
	XGBoost	0.0000	0.8872	0.5221
	LogisticRegression	0.0958	0.8948	0.5380
	SVM	0.1406	0.9036	0.6247
Tree_Nuts	RandomForest	0.2604	0.6982	0.6592
	XGBoost	0.4378	0.7404	0.7172
	LogisticRegression	0.3754	0.7029	0.6821
	SVM	0.4414	0.7190	0.6296

TABLE 3.11 – Performance des modèles pour les sous-types d’allergies alimentaires (ALEX).

Chapitre 4

Analyse et Évaluation Des Modèles

4.1 Choix du modèle..

Pourquoi ai-je choisi le F1-score de la classe 1 ?

Tout d’abord, rappelons que le F1-score est défini comme suit :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

avec :

$$\text{Précision} = \frac{TP}{TP + FP}, \quad \text{Rappel} = \frac{TP}{TP + FN}$$

où TP = Vrais Positifs, FP = Faux Positifs et FN = Faux Négatifs.

Notre projet consiste à classifier les cas d’allergie en deux classes : 0 (pas d’allergie) et 1 (présence d’allergie). L’objectif principal est de détecter si une personne est allergique. Dans ce contexte, il est plus pertinent de se concentrer sur le F1-score de la classe 1, car il évalue la capacité du modèle à identifier correctement les personnes allergiques. À l’inverse, le F1-score de la classe 0 est moins critique : il est moins problématique de classer par erreur une personne non allergique comme allergique (faux positif) que de ne pas détecter une allergie réelle (faux négatif), ce qui pourrait avoir des conséquences graves.

C’est pourquoi, dans notre démarche de sélection de modèle, nous avons choisi de nous baser principalement sur le F1-score de la classe 1.

Le modèle **XGBoost** a également montré une excellente stabilité lors des différents tests réalisés (validation croisée et tests finaux). Sur les trois jeux de tests, il a systématiquement produit des matrices de confusion cohérentes, présentant toujours un très faible nombre de faux négatifs et une proportion plus élevée de

faux positifs, ce qui est acceptable dans le cadre de notre problématique. Cette robustesse confirme la fiabilité de XGBoost pour la détection des cas d'allergie.

Voici un exemple de matrice de confusion obtenue avec XGBoost :

R��el	Pr��dit	
	Non allergique (0)	Allergique (1)
Non allergique (0)	1121	0
Allergique (1)	1	1228

TABLE 4.1 – Matrice de confusion du mod  le XGBoost pour la classification Allergie/Non-Allergie (**V1**)

Prenons par exemple la cible **Severe_Allergy** du test V1. Nous pouvons visualiser la courbe AUC-ROC du mod  le **XGBoost** appliqu      cette cible. Cette courbe permet d'  valuer la capacit   du mod  le    distinguer correctement les personnes allergiques s  v  res des non-allergiques    diff  rents seuils de d  cision. Comme le montre la figure ci-dessous, le mod  le XGBoost obtient une courbe AUC-ROC proche de l'optimum, traduisant une excellente capacit   de discrimination.

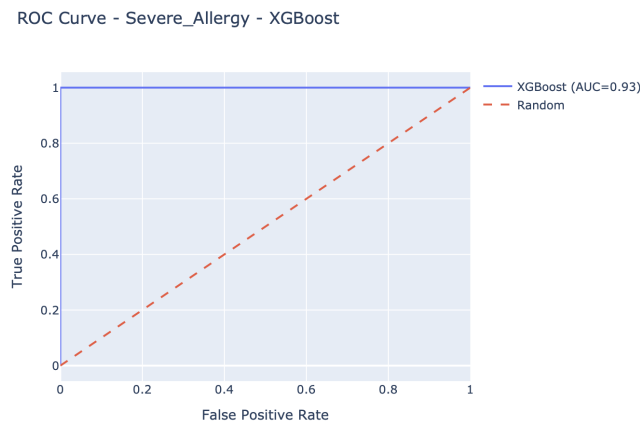


FIGURE 4.1 – Courbe AUC-ROC du mod  le XGBoost.

En revanche, certains autres mod  les tels que la r  gression logistique et le SVM ont produit des courbes AUC-ROC moins satisfaisantes. Ces courbes montrent une capacit   r  duite    discriminer correctement les personnes allergiques des non-allergiques, ce qui traduit une performance globale plus faible.

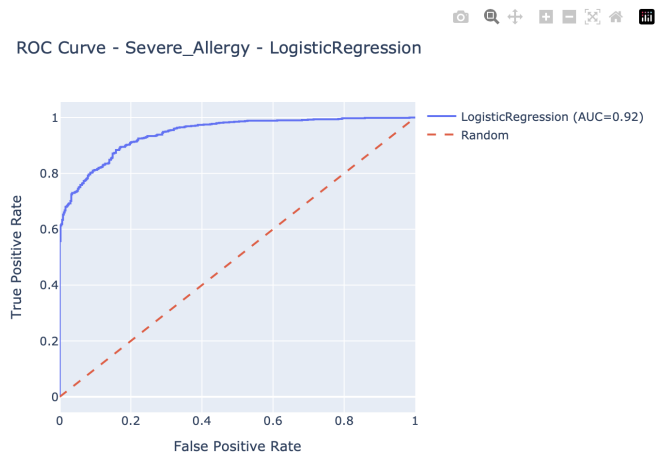


FIGURE 4.2 – Courbe AUC-ROC du modèle Régression Logistique.

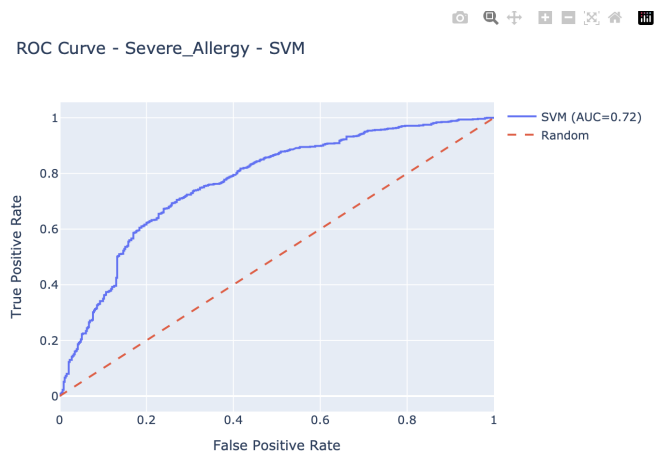


FIGURE 4.3 – Courbe AUC-ROC du modèle SVM.

Ainsi, ces résultats confirment la supériorité du modèle XGBoost dans notre contexte de détection des cas d'allergie.

Il est important de noter que cette supériorité ne se limite pas uniquement à la cible Severe_Allergy, mais se vérifie également sur l'ensemble des autres cibles du projet.

Tous les résultats détaillés — incluant les F1-scores pondérés, les précisions, les matrices de confusion, et les courbes AUC-ROC — ont été sauvegardés dans les rapports générés par les notebooks et dans des fichiers CSV, afin d'assurer la traçabilité et la comparaison entre les différents modèles et Tests.

4.2 Analyse des variables les plus contributives à la classification

Une analyse des *features* les plus influentes a été réalisée pour chaque test, en utilisant les importances calculées par le modèle **XGBoost**. Cette analyse permet de mieux comprendre quelles variables ont le plus contribué à la détection des cas d'allergie dans notre contexte.

Voici quelques exemples qui ont particulièrement attiré notre attention :

Pour le test **V1** et la cible **Severe_Allergy**, les variables les plus importantes sont identifiées dans la figure ci-dessous. On peut remarquer que la sévérité de l'allergie est bien capturée par le modèle, notamment grâce aux variables relatives à la sévérité de l'ARIA (codée de 1 à 5, ce qui reflète le degré de gravité de l'allergie), ainsi qu'aux symptômes oraux et cutanés, qui apparaissent comme des facteurs fortement discriminants.

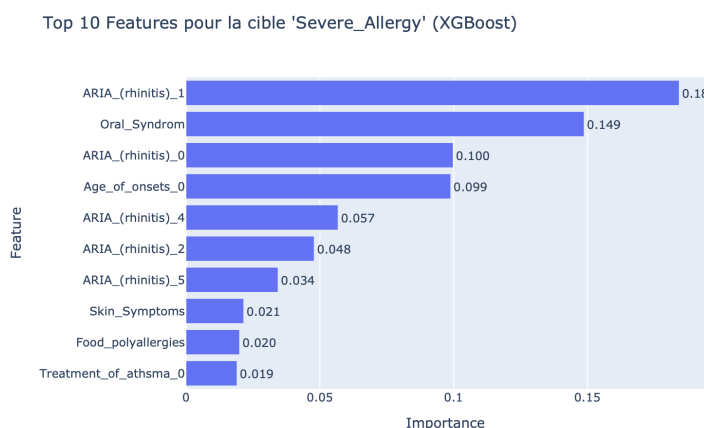


FIGURE 4.4 – Variables les plus importantes pour la classification des cas **Severe_Allergy** dans le test V1.

Un deuxième exemple concerne la cible **Venom_Allergy** avec le même test. On remarque que la variable qui contribue le plus à la classification est issue des colonnes créées à partir des allergènes de venin. Cela montre que les allergies au venin sont principalement détectées à partir des tests sanguins (IgE spécifiques) et non à partir des symptômes cliniques, qui semblent moins informatifs dans ce contexte.

Top 10 Features pour la cible 'Venom_Allergy' (XGBoost)

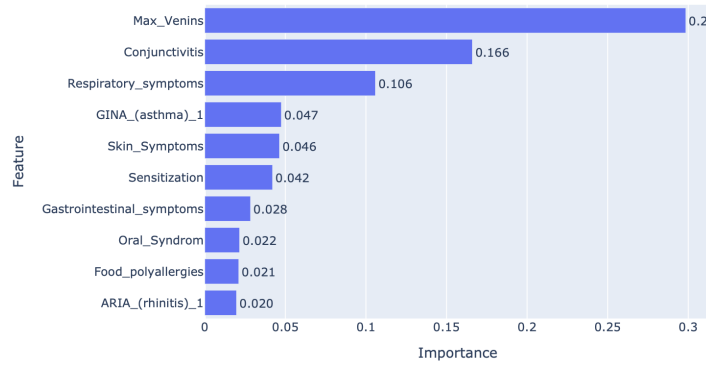


FIGURE 4.5 – Variables les plus importantes pour la classification des cas **Venom_Allergy** dans le test V1.

Un autre exemple est celui de la classification des cas d'allergie de pollen à partir du test ALEX. Les variables les plus importantes identifiées incluent des symptômes respiratoires, ainsi qu'une variable **general_cofactor_1** qui indique la présence d'un facteur général lié à une activité physique ou à un effort, ce qui peut être un élément déclencheur d'une réaction allergique. On note également l'importance des allergènes spécifiques de pollen et d'arbres mesurés lors du test, qui apparaissent comme des facteurs discriminants majeurs dans la détection de cette allergie.

Top 10 Features pour la cible 'Type_of_Respiratory_Allergy_IGE_Pollen_Herb' (XGBoost)

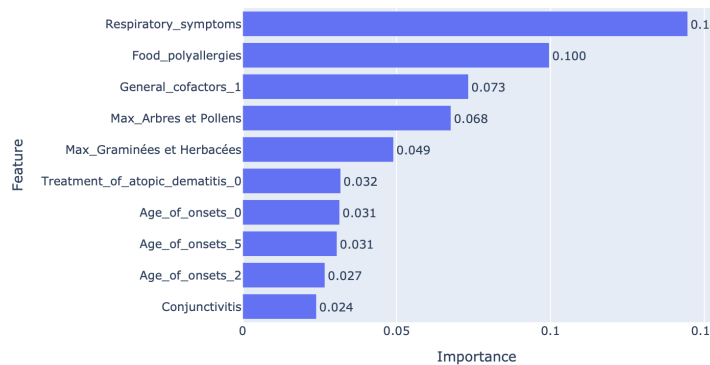


FIGURE 4.6 – Variables les plus importantes pour la classification des cas d'allergie au pollen dans le test ALEX.

Enfin, pour la cible **Food_Allergy** dans le test ALEX, on observe que les symptômes gastro-intestinaux ont joué un rôle principal dans la classification. Les allergènes issus des légumes apparaissent également comme des contributeurs majeurs, accompagnés de symptômes oraux et respiratoires qui renforcent la capacité du modèle à détecter cette allergie alimentaire.

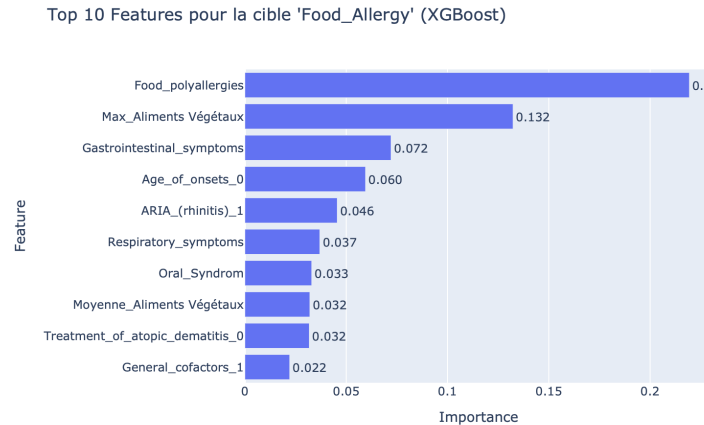


FIGURE 4.7 – Variables les plus importantes pour la classification des cas Food_Allergy dans le test ALEX.

4.3 Hyperparamètres

J'ai également pris en charge la recherche et l'optimisation des hyperparamètres du modèle **XGBoost** afin d'améliorer ses performances pour chaque test (ISAC_V1, ISAC_V2, ALEX). Ces hyperparamètres ont été déterminés à l'aide d'une recherche systématique et enregistrés dans des fichiers dédiés, permettant une traçabilité complète et une réutilisation facile pour les futures analyses.

Chapitre 5

Conclusion

Ce projet a permis de mettre en lumière la capacité des modèles d'apprentissage automatique, et en particulier du modèle **XGBoost**, à classifier de manière efficace et robuste les différents types d'allergies à partir de données complexes issues de puces allergéniques. Grâce à une méthodologie rigoureuse incluant l'exploration des données, l'équilibrage des classes, et l'évaluation des performances par validation croisée, nous avons pu comparer plusieurs modèles (Random Forest, XGBoost, Régression Logistique et SVM) sur différents jeux de tests (ISAC_V1, ISAC_V2 et ALEX) et pour différentes cibles.

Les résultats obtenus ont clairement démontré la supériorité du modèle XGBoost, notamment sur les métriques clés telles que le F1-score, l'AUC-ROC, et la précision, en particulier pour la détection des cas d'allergie sévère. En complément, l'analyse des variables les plus contributives a permis d'identifier des facteurs d'importance majeure dans la classification, offrant des perspectives intéressantes pour une meilleure interprétation biologique des résultats.

Enfin, l'ensemble des résultats détaillés, incluant les métriques d'évaluation, les matrices de confusion, les courbes AUC-ROC et l'importance des variables, ont été soigneusement documentés dans les notebooks et les fichiers CSV, garantissant une traçabilité complète du travail réalisé.

Ces travaux ouvrent la voie à de futures améliorations, notamment en explorant d'autres approches d'explicabilité (SHAP, LIME), en affinant l'ingénierie des features, et en intégrant des données cliniques supplémentaires pour enrichir le modèle. Ils illustrent également l'intérêt des méthodes d'intelligence artificielle pour soutenir le diagnostic en allergologie et contribuent ainsi à une meilleure compréhension des mécanismes allergiques.

Chapitre 6

Bibliographie

- <https://www.data.gouv.fr/en/datasets/allergen-chip-challenge/>
- <https://www.health-data-hub.fr/actualites/publication-base-de-donnees-allerge>
- <https://sfa.lesallergies.fr/articles-evenements/lallergen-chip-challenge-resu>
- <https://github.com/Trustii-team/AllergenChip>