GESTION DES TACHES COUPLEES INVENTIVES PAR PROJETS SUCCESSIFS PAR EXTENSION DES CRITERES DE PERFORMANCE

RESUME DU CHAPITRE 9

Ce chapitre a pour ambition d'analyser les effets de la capitalisation sur le challenge 2 du programme Epidemium. Dans un premier temps, nous analyserons le déroulement du challenge 2 et nous présenterons les résultats obtenus. Nous montrerons que les bonnes performances du challenge 2 sont principalement dues à la capitalisation menée par les participants eux-mêmes. Nous mettrons ensuite en avant un effet de double extension qui ne pouvait être observé sur un seul challenge. D'abord une extension dans la taille des espaces explorés avec l'apparition de projets qui ne proviennent pas des bases de données disponibles. Ensuite une extension des fonctions de valeur via des projets qui ne peuvent être évalué uniquement avec les critères de valeurs qui ont été créés durant le challenge 1. Nous suggérons que ces phénomènes d'extension doivent être gérés durant le processus de capitalisation.

Nous avons vu dans le chapitre 8 que la capitalisation séquentielle dans le cas de tâches couplées inventives impliquait de piloter la transmission entre les projets de l'avancement de l'exploration des espaces. Dans ce chapitre, nous étudions les effets de la capitalisation qui a été mise en œuvre par les organisateurs d'Epidemium sur la performance du deuxième challenge, et nous tirons des conclusions sur les moyens à mettre en œuvre pour piloter la capitalisation séquentielle.

1. BILAN GLOBAL DU DEUXIEME CHALLENGE

1.1. UNE PLUS GRANDE COMMUNAUTE AVEC MOINS DE PARTICIPATION

Entre la fin du premier et du deuxième challenge la communauté Epidemium, définie comme le nombre de personnes inscrites aux meetups et aux newsletter, a doublé passant de 600 à 1200 inscrits. En parallèle, le nombre de participants actifs a diminué sur le deuxième challenge : 22 volontaires ont participé sur les deux premières thématiques tandis que le thème 3 a regroupé 32 étudiants issus des écoles Centrale Paris, Polytechnique et ESIEA. Cela représente un total de 54 membres (dont 32 étudiants), soit un taux de transformation de 4.5% (2% sans les étudiants). Ce taux est beaucoup plus faible que durant le premier challenge, où il y a eu environ 10% de la communauté ayant été active. Les raisons sont multiples. Les participants évoquent des barrières à l'entrée plus élevée que durant le premier challenge rendant la participation plus complexe. En effet, les organisateurs ont imposé la cohérence du projet avec la littérature scientifique. De ce fait, les conditions pour soumettre une proposition ont demandé un travail plus structuré et plus approfondi pour les participants. Deuxièmement, certains participants ont reconnu qu'il y avait une double difficulté d'acculturation pour intégrer le programme Epidemium : d'abord pour comprendre le monde relatif au domaine médical et scientifique, d'autant plus renforcée par l'obligation de fournir un résultat scientifique. Ensuite la mise en pratique d'une culture de l'ouverture et de la collaboration qu'on ne retrouve pas traditionnellement dans les projets scientifiques. Enfin, le projet Epidemium ne bénéficie plus de l'effet de nouveauté.

1.2. PRESENTATION DES PROJETS

13 équipes ont proposé un projet, dont 10 sont arrivés jusqu'au bout en soumettant leur projet au comité à la fin du challenge. Sur les 10 projets, 5 sont des projets d'étudiants relatifs à la thématique 3. Les 10 projets diffèrent dans leur approche suivant les thématiques auxquelles ils sont raccrochés (**tableau 19**). La thématique 1 ne contient qu'un seul projet, *IDEA*, dont l'objectif est de développer un outil de visualisation des données basé sur un algorithme de machine learning. Le principe est qu'à chaque fois que l'outil sera utilisé, la visualisation qu'il aura choisie (histogramme, carte, courbes,…) sera notée par l'utilisateur en fonction de plusieurs paramètres comme le type d'utilisateur ou le type de base de données. Ainsi plus l'outil est utilisé, plus il « apprend » de ses utilisateurs afin de fournir le type de visualisation le plus adapté. Ce projet est réalisé par des employés de l'entreprise CONIX, qui avaient déjà participé au premier

challenge dans le cadre du projet ELSE. Une deuxième catégorie d'équipes issus de la thématique 2 a cherché à développer des outils algorithmiques pour prédire la progression du cancer dans le temps et dans l'espace en fonction du régime alimentaire (*Cancer Diet*), pour étudier l'impact des essais cliniques (*AROUND*) ou encore pour aider les autorités publiques à prendre des décisions en terme de santé publique (*Locapred*). Une troisième catégorie concerne les projets étudiants sur la prédiction de la mortalité des cancers dans les pays en voie de développement (*Prévenir pour mieux guérir, Cancerinfl, Osy3a, Oma, Octopus*). Contrairement aux thématiques proposées durant le challenge 1, la thématique 3 se concentre sur un nombre de variables limitées et cherche à multiplier les explorations dans un espace restreint de l'espace des hypothèses. Pour rappel, la thématique 3 est de *prédire la mortalité par cancer dans les pays en voie de développement*. Dans cette thématique, les organisateurs précisent le type d'impact {mortalité}, demandent à ne travailler que sur un seul type de cancer, et réduisent l'exploration aux {pays en voie de développement}.

Enfin, un projet unique est issu d'une association appelée *Cancer Au Travail (CAT)*, qui s'intéresse à un champ encore peu exploré par la littérature scientifique : l'impact du modèle social d'un pays, sa vitalité démocratique, ses conditions de travail ou son modèle de production et leur impact sur la survivance du citoyen ? Cette notion de survivance fait émerger un concept qui n'existe pas dans la littérature médicale ni dans les données Epidemium, et qui fait référence à l'étude des patients guéris du cancer. Habituellement, la littérature médicale s'intéresse à cette population quand les scientifiques cherchent à quantifier les taux de survie au cancer et étudier l'efficacité d'un traitement. Ici, le projet CAT propose d'étudier le patient dans son environnement social, et l'impact que cet environnement peut avoir sur sa guérison ou sa rechute. Cela se modélise dans l'espace des hypothèses comme une extension de l'espace par l'intégration d'une nouvelle relation (**figure 55**), définie comme {survivance}. L'exploration ici n'est donc plus limitée aux seules données disponibles ou existantes, mais peut s'étendre par l'apparition de nouveaux mots.

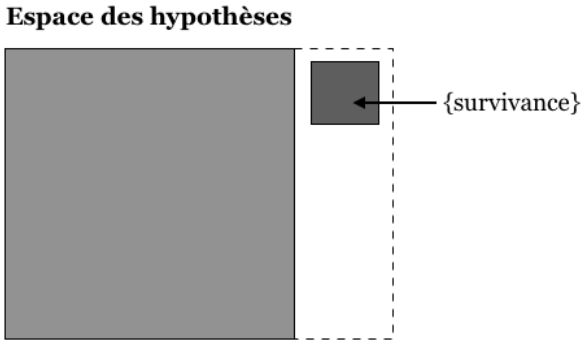

Figure 55. Extension de l'espace des hypothèses par le projet CAT et la notion de « survivance ».

Thème 1 - Construire une visualisation de données de l'incidence des cancers en exposant les facteurs épidémiologiques associés à leur dynamique		
IDEA	Développer un outil de visualisation des données sur le cancer qui va apprendre la bonne représentation en fonction des bases de données, du type d'utilisateur,... pour proposer la solution la plus adaptée	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
Thème 2 - Développer un outil prédictif pour la progression du cancer dans le temps et dans l'espace, en fonction des facteurs connus ou supposés qui déterminent son évolution.		
Cancer Au Travail	Déterminer le niveau de survivance du cancer vis-à-vis des facteurs de risques sociétaux (type de sécurité sociale, travail,...)	$\mathcal{A}$ = *{impact =* ***survivance****} ({type de cancer}, {facteurs de risque =* ***sociétaux****})*
Locapred	Construire un outil prédictif afin d'aider les autorités à prendre des décisions en terme de santé publique	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
Cancer Diet	Développer un algorithme pour prédire le taux de mortalité du cancer en fonction du régime alimentaire	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer}, {facteurs de risque =* ***régime alimentaire****})*
AROUND	Déterminer l'impact des essais cliniques sur l'incidence et la mortalité des cancers	$\mathcal{A}$ = *{impact =* ***incidence, mortalité****} ({type de cancer}, {facteurs de risque =* ***essais cliniques****})*
Thème 3 – Prédire dans le temps et dans l'espace la mortalité des cancers dans les pays en voie de développement		
Prévenir pour mieux guérir	Prédire les taux de mortalité des cancers digestifs dans des pays dont les régimes alimentaires et les conditions environnementales sont différentes : la France et le Brésil	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestifs****}, {zone géographique =* ***France, Brésil****})*
Cancerinfl	Etude de l'évolution de la mortalité des cancers gynécologiques dans les pays d'Asie en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***gynécologiques****}, {zone géographique =* ***pays d'Asie en développement****})*
Osy3A	Prédire l'évolution de la mortalité des cancers digestifs dans les pays en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestif****}, {zone géographique =* ***pays en développement****})*
Oma	Etude de l'évolution de la mortalité du cancer de l'estomac dans les pays en voie de développement	$\mathcal{H}$ = *{impact =* ***mortalité****} ({type de cancer =* ***estomac****}, {zone géographique =* ***pays en développement****})*
Octopus	Prédire l'évolution de la mortalité des cancers colorectaux dans les pays en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***colorectaux****}, {zone géographique =* ***pays en développement****})*

Note : les catégories spécifiées par les projets sont marquées en gras

Tableau 19. Présentation des projets du challenge 2 d'Epidemium.

1.3. BILAN DU DEUXIEME CHALLENGE4CANCER

Bien que la communauté Epidemium ait doublée entre le premier et le deuxième challenge, le nombre de participants actifs (hormis les équipes d'étudiants) a fortement diminué. La hausse de la barrière à l'entrée exigée par les organisateurs (revue de littérature et validité scientifique) a certainement contribué à la baisse des incitations des participants. En contrepartie, les projets soumis à la fin du challenge sont de bien meilleure qualité : dans l'ensemble, tous les outils développés par les participants sont beaucoup plus aboutis, et certains d'entre eux ont mené à une publication dans des revues scientifiques. De même, les hypothèses formulées se rapprochent beaucoup plus des exigences de validité d'hypothèses scientifiques en épidémiologie.

Notre analyse du challenge dans ce chapitre montrera que les organisateurs ont mieux répondu à certains besoins en terme de gestion durant le challenge, notamment en mettant à disposition des participants une équipe de scientifiques spécialistes en épidémiologie et capables d'aider et d'aiguiller la production. De plus, nous montrerons que la réussite du challenge est également dû à une capitalisation réalisée par les organisateurs entre le challenge 1 et le challenge 2 sur les pistes intéressantes à explorer à partir des données disponibles. En effet, la seule présence des bases de données ne permet pas aux organisateurs d'anticiper quelles sont les bonnes explorations à mener et donc de formuler des thématiques pertinentes. Au contraire, le processus d'exploration pour la tâche couplée inventive demande une gestion par la mise en place de challenges successifs. C'est un processus de **capitalisation séquentielle** : d'abord les participants explorent les bases de données durant les challenges afin de reformuler et de préciser quelles sont les bonnes hypothèses à explorer. Dans un deuxième temps, les organisateurs capitalisent sur la production globale entre les challenges afin d'orienter les explorations futures.

La présentation de ce deuxième challenge reprendra celle que nous avons réalisée pour le premier. Nous présenterons d'abord le processus d'exploration mené par les équipes. Nous nous intéresserons particulièrement à un nouveau processus d'exploration mené par l'équipe CAT. Nous analyserons ensuite la production globale du challenge au travers des métriques que nous avons présenté dans le chapitre 8. Enfin, nous proposerons une analyse critique de la capitalisation durant le challenge.

2. *EXPLORATION DES ESPACES ET EVALUATION DE LA PRODUCTION*

2.1. LE PROJET CAT COMME EXTENSION DE L'ESPACE DES HYPOTHESES

De manière générale, les projets orientés vers la formulation d'hypothèses scientifiques basées sur les données ont suivi le même processus d'exploration que celui présenté lors du premier challenge. Après avoir formulé leur hypothèse de départ à partir des données disponibles, les équipes ont confronté l'hypothèse aux données disponibles et ont procédé à une reformulation cohérente avec les données et les analyses fournies. Si les projets ont suivi peu ou prou le même processus que durant le premier challenge, le projet CAT nécessite une analyse distincte de par sa spécificité. En effet, contrairement aux méthodes déployées par les autres participants, le projet est initié non pas à partir des bases de données disponibles mais par une interrogation que les porteurs de projet ont sur le cancer et la notion de survivance. Les premiers travaux ont consisté à explorer à la fois les bases de données disponibles et la littérature pour situer le projet. Le projet CAT a notamment été soutenu et accompagné par la communauté de scientifiques mises à disposition par les organisateurs d'Epidemium et les laboratoires Roche. Suite à leur première exploration des données, un premier constat apparaît : la catégorie de publication "survie et modèle social" n'existe pas dans les bases étudiées et le sujet est principalement traité dans la littérature sous l'angle des cancers professionnels et les questions du travail sont abordées dans le cadre de la psycho-oncologie sous l'angle des "traces psychique" ou "la fabrique psychique du cancer".

En conséquence, les bases de données récoltées ne sont pas suffisantes pour rendre compte du phénomène. De plus, le projet CAT réalise rapidement que les quelques bases de données qui pourraient être utiles sont de mauvaise qualité. Grâce aux meetups, l'équipe projet se met en relation avec le projet Locapred qu'il convainc de nettoyer les bases de données OIT (Organisation Internationale du Travail). Cette organisation fournit un ensemble de données relatives au travail en fonction des pays. Cette collaboration permet d'aligner les objectifs des deux projets : d'un côté, le projet CAT va obtenir des bases de données fiables pour son hypothèse, tandis que le projet Locapred va utiliser les bases de données OIT comme matière première pour consolider son algorithme. Ces bases de données ont permis au projet CAT de constituer 5 familles de métriques pour décrire le modèle social en lien avec les incidences des cancers (tous les types de cancers) :

- Indicateurs du collectif public (conventions collectives, pourcentage de population avec ou sans couverture sociale, accessibilité au soin et aux soignants, dépenses de santé)
- Dynamisme social (niveau de productivité, de formation et de revenus, taux de chômage et de pauvreté)
- Production dominante/type d'activité (prédominance agricole / industrielle /tertiaire, protection sociale, médecine du travail)

- Conditions de travail (horaires, travail des enfants, économie déclarée ou non,…)
- Indicateurs socio-démographiques (disparité hommes/femmes, présence des seniors au travail,…)

Ainsi, au lieu d'adapter leur hypothèse aux bases de données existantes, le projet CAT a choisi la stratégie opposée. L'équipe de participants a cherché à définir les différents éléments de langage de son hypothèse ({travail}, {survivance}) au travers de variables déjà existantes dans les bases de données. On comprend qu'il y a projection de la question de recherche sur les données déjà existantes (**figure 56**). Chaque terme utilisé dans son hypothèse est retranscrit au travers d'un ensemble de termes déjà existants dans les bases de données. La majeure partie du travail consiste à retranscrire de la manière la plus fidèle possible les concepts développés au début du projet.

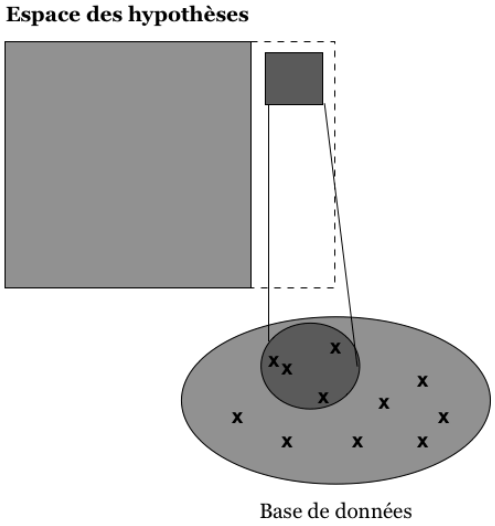

Figure 56. Projection de l'hypothèse de recherche sur les bases de données existantes.

Suite à ces premières analyses, le projet CAT est entré en relation avec l'institut Curie et l'institut Gustave Roussy pour définir des stratégies possibles afin de continuer le projet. En effet, ces instituts ont considéré comme intéressants l'approche de CAT, et étaient prêts à aller chercher plus loin.

Le processus d'exploration suivi par le projet Cancer Au Travail diffère des processus suivis par les autres équipes :

- *Formulation d'une hypothèse de travail indépendamment des données* : l'hypothèse initiale est formulée à partir des connaissances existantes en épidémiologie et sur les observations réalisées. En ce sens , la formulation correspond à une approche knowledge-driven

- *Projection de l'hypothèse sur les bases de données existantes* : L'équipe sélectionne des catégories issues des bases de données susceptibles de correspondre aux catégories de l'hypothèse de travail. Il y a projection de l'hypothèse sur les bases de données
- *Nettoyage, traitement et collecte de données* : Une fois les hypothèses choisies et les données déterminées, les participants nettoient les bases de données
- *Exploration des bases de données* : L'équipe développe un algorithme pour analyser les données et vérifier la valeur de vérité de l'hypothèse.
- *Processus d'optimisation* : une fois que l'hypothèse découverte est valide d'un point de vue de la communauté scientifique, les équipes projets cherchent à optimiser le modèle algorithmique utilisé pour analyser les bases de données. Aucune équipe projet n'a abouti à cette étape durant le premier challenge, probablement par manque de temps.

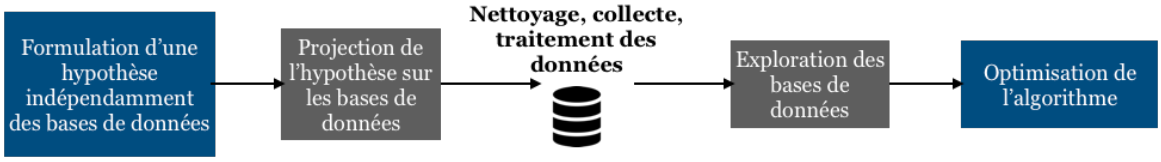

Figure 57. Processus d'exploration durant le premier challenge

2.2. EVALUATION FINALE DES PROJETS PAR LES COMITES

Les projets ont été évalués en deux temps par le comité scientifique : les organisateurs ont séparé l'évaluation des projets de la thématique 3 qui a été conçue spécifiquement pour des projets étudiants. De plus, les projets des thématiques 1 et 2 se sont déroulés entre Juin 2017 et Janvier 2018 alors que les projets de la thématique 3 se sont déroulés entre Novembre 2017 et Mars 2018. A l'aide de critères similaires à ceux utilisés dans le challenge 1, trois projets lauréats ont été choisis pour les thématiques 1 et 2 (dans l'ordre décroissant) : Locapred, IDEA et Cancer Diet. Bien que leurs approches soient originales d'un point de vue scientifique, les projets CAT et AROUND n'ont pas réussi à fournir à la fin du challenge des rendus suffisamment avancés par rapport aux autres projets. Cependant, ils ont été récompensés par leur côté éthique et leur originalité dans l'approche envisagée. Contrairement au premier challenge, le classement s'est avéré plus sélectif et les organisateurs ont souhaité mettre en avant les projets les plus aboutis.

Concernant la thématique 3, les projets ont été évalués indépendamment dans le cadre scolaire, et ont également reçu une évaluation par le comité scientifique. Les trois projets lauréats sont dans l'ordre décroissant : Osy3A, Octopus, Oma. Mis à part le projet vainqueur Osy3A, le résultat est en demi-teinte selon le comité. En effet, les étudiants semblaient ne pas avoir pris conscience de l'importance que le projet soit cohérent avec le cadre de recherche habituel dans la littérature en

épidémiologie. La plupart des projets ont proposé des modèles prédictifs agrégés, réduisant l'intérêt scientifique et en terme de politique publique pour l'exploitation des résultats.

2.3. ANALYSE DE LA PRODUCTION ET EXTENSION DE LA FONCTION DE VALEUR

Nous avons de notre côté évalué la production au travers des mêmes métriques que celles utilisées lors du challenge 1.

2.3.1. Formulation et vérification des hypothèses

Malgré un plus faible nombre de projets que dans le premier challenge, le nombre d'hypothèses générées durant le challenge est important, avec 12 hypothèses dont 8 axes de travail et 4 hypothèses canoniques. Ce résultat vient principalement du grand nombre d'hypothèses formulées par les étudiants via la problématique du cancer dans les pays en voie de développement. Pour autant ces hypothèses sont essentiellement des axes de travail qui étudient plusieurs cancers en même temps, et il n'a pas été possible d'extraire d'hypothèses scientifiquement intéressantes de leurs travaux. Ce problème avait déjà été identifié durant le premier challenge comme l'illusion des données. En effet, les communications entre étudiants et communauté scientifique ont été assez rare, et les scientifiques n'ont pas pu orienter les travaux afin d'éviter ce type d'exploration peu fécond.

Alors qu'aucune hypothèse formulée par les étudiants n'a été transformée en hypothèse canonique vérifiée, leurs travaux ont permis en revanche de mieux comprendre le fonctionnement des algorithmes de machine learning dans la recherche de corrélation entre facteurs de risque et incidence ou mortalité d'une pathologie. La plupart des projets étudiants ont permis d'extraire au travers de leurs modèles prédictifs une liste des facteurs de risque les plus discriminants pour une pathologie en fonction de son « poids » dans le modèle algorithmique. A partir de ces listes, les scientifiques pourraient se concentrer sur des facteurs de risque ou des facteurs protecteurs qui semblent jouer un rôle fondamental dans le développement de la pathologie. A titre d'exemple le projet Locapred a testé son modèle prédictif sur l'incidence du cancer du colon par genre et groupe d'âge. A partir de leur analyse, ils ont pu extraire les quatre facteurs les plus importants pour estimer l'incidence du cancer du colon : le taux de pollution de l'air à 2.5PPM, la consommation de cigarette, la consommation domestique de fruits, le taux de chômage. Sans apporter de résultat scientifique probant, cette analyse permet de formuler des hypothèses canoniques sur des facteurs de risques et protecteurs soupçonnés pour le cancer du colon. De plus, leur analyse met en avant la valeur potentielle des bases de données utilisées pour tester l'hypothèse.

Le projet CAT n'a pas pu fournir d'hypothèses canoniques à la fin du challenge. Le projet est en effet dans une phase très amont et leurs travaux ont permis pour l'instant d'identifier quelques

variables potentiellement utiles pour étudier quantitativement leur hypothèse. Enfin, les projets Cancer Diet et AROUND ont réalisé des explorations intéressantes mais ne possédaient pas les compétences suffisantes en terme d'analyse de données pour obtenir des résultats convaincants d'un point de vue scientifique. Il est à noter cependant que l'exploration du projet AROUND a permis de formuler une hypothèse canonique originale et potentiellement viable d'un point de vue scientifique.

Projets	**Hypothèse**	**Analyse par le langage**	**Bases de données**
Cancer Au Travail	Déterminer le niveau de survivance du cancer vis-à-vis des facteurs de risques sociétaux (type de sécurité sociale, travail,…)	$\mathcal{A}1$ = *{impact =* ***survivance****} ({type de cancer}, {facteurs de risque =* ***sociétaux****}*	Bases de données ILO
Cancer Diet	Développer un algorithme pour prédire le taux de mortalité du cancer en fonction du régime alimentaire	$\mathcal{A}2$ = *{impact =* ***mortalité****} ({type de cancer}, {facteurs de risque =* ***régime alimentaire****})*	Données FAO
AROUND	Déterminer l'impact des essais cliniques sur l'incidence et la mortalité des cancers	$\mathcal{A}3$ = *{impact =* ***incidence, mortalité****} ({type de cancer}, {facteurs de risque =* ***essais cliniques****})*	Essais cliniques (clinicaltrials.org), données INSEE (mortalité cancer en France)
AROUND	Déterminer l'impact des essais cliniques sur la mortalité des cancers des poumons, du pancréas, du sein et colorectal dans 5 régions françaises	$\mathcal{A}4$ = *{impact =* ***mortalité****} ({type de cancer =* ***poumon, pancréas, sein colorectal****}, {facteurs de risque =* ***essais cliniques****}, {zone géographique =* ***5 régions françaises****})*	Essais cliniques (clinicaltrials.org), données INSEE (mortalité cancer en France)
Prévenir pour mieux guérir	Prédire les taux de mortalité des cancers digestifs dans des pays dont les régimes alimentaires et les conditions environnementales sont différentes : la France et le Brésil	$\mathcal{A}5$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestifs****}, ({facteur de risque =* ***régime alimentaire****}, {zone géographique =* ***France, Brésil****})*	Données mortalité, FAO, World Bank data
Cancerinfl	Etude de l'évolution de la mortalité des cancers gynécologiques dans les pays d'Asie en voie de développement	$\mathcal{A}6$ = *{impact =* ***mortalité****} ({type de cancer =* ***gynécologiques****}, {zone géographique =* ***pays d'Asie en développement****})*	World Bank data, FAO, ILOstat
Osy3A	Prédire l'évolution de la mortalité des cancers digestifs dans les pays en voie de développement	$\mathcal{A}7$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestif****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO
Octopus	Prédire l'évolution de la mortalité des cancers colorectaux dans les pays en voie de développement	$\mathcal{A}8$ = *{impact =* ***mortalité****} ({type de cancer =* ***colorectaux****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO
Locapred	Les particules 2.5PPM dans l'air sont un facteur de risque de l'incidence dans le cadre des cancers du colon	$\mathcal{H}1$ = *{impact =* ***incidence****} ({type de cancer =* ***colon****}, {FdR =* ***polluation air 2.5PPM****}, {profil =* ***féminin, âge 20-29 ans****})*	Données mortalité, FAO, World Bank data

Locapred	La consommation de fruit est un facteur de présention de l'incidence dans le cadre des cancers du colon	$\mathcal{H}1$ = *{impact =* ***incidence****} ({type de cancer =* ***colon****}, {FdR =* ***consommation de fruits****}, {profil =* ***féminin, âge 20-29 ans****})*	Données mortalité, FAO, World Bank data
Oma	Etude de l'évolution de la mortalité du cancer de l'estomac dans les pays en voie de développement	$\mathcal{H}3$ = *{impact =* ***mortalité****} ({type de cancer =* ***estomac****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO

Tableau 20. Récapitulatif des hypothèses formulées durant le challenge 2

2.3.2. Elaboration d'outils pour l'exploration dans les espaces

Contrairement au premier challenge, les projets d'outils sont plus aboutis. Les deux équipes (IDEA et Locapred) ayant choisi de construire des outils pour l'exploration des données sont partiellement issues d'équipes ayant participé au premier challenge, et elles avaient anticipé certains problèmes et points durs à éviter. Le premier challenge leur a permis un premier contact avec les données qui les a mené à identifier les points faibles de l'exploration. Le projet IDEA a permis de fournir à la fin du challenge à un prototype fonctionnel de son logiciel de visualisation des données. Les scientifiques y voient un réel potentiel pour la recherche en épidémiologie c'est pourquoi l'équipe a été récompensée en étant en seconde place. Depuis la fin du challenge, l'équipe IDEA multiplie les tests de l'algorithme dans des cercles médicaux afin d'améliorer la qualité de l'outil. En fonction des résultats obtenus, l'équipe IDEA a pour objectif de soumettre une publication scientifique sur leur méthode originale.

Le projet Locapred, grand gagnant du challenge 2, a quant à lui développé un outil servant à nettoyer les bases de données ouvertes contenant des éléments manquants ou aberrants. En effet, une des conclusions majeures de la fin du premier challenge était la difficulté d'explorer les bases de données étant donné leur qualité souvent médiocre. Cet outil est utilisable clé en main pour des personnes non expertes, mais son code est également librement disponible pour des spécialistes de l'analyse de données afin de modifier tout ou partie des paramètres choisis. Les premières bases de données constituées par le logiciel ont intéressé une doctorante qui travaille en Angleterre sur des méthodes de nettoyage des bases de données. Leur collaboration a permis de publier un papier dans une conférence scientifique sur le machine learning (Chelly Dagdia et al., 2018). De plus, les membres du projet continuent de travailler sur le logiciel dont la partie sur la partie données manquantes va faire l'objet de deux publications : une dans le domaine de l'épidémiologie sur les questions des données ouvertes, et une technique sur la question des méthodes de remplissage des données.

2.3.3. Extension de la fonction de valeur

Le challenge 2 a enregistré une baisse importante de la participation (en dehors des projets étudiants) par rapport au challenge 1. Cependant, les projets menés ont été beaucoup plus aboutis et ont mené pour certains à des publications scientifiques à la fois dans le domaine de l'épidémiologie mais également dans le domaine du machine learning. De plus, les différents instituts partenaires du programme Epidemium ont engagé des collaborations avec au moins trois des projets finaux (CAT, Locapred, IDEA). Ainsi, contrairement au premier challenge, certaines équipes continuent leur projet hors du cadre d'Epidemium. Si les résultats sont assez positifs pour les projets des thèmes 1 et 2, le résultat est plus en demi-teinte pour les projets étudiants. En effet, les organisateurs n'ont pas cherché à tirer suffisamment parti de la présence d'un nombre important de participants pour piloter l'exploration autour des questions du cancer dans les pays en voie de développement. Ainsi, les projets des étudiants n'ont pas permis d'apporter un nouveau regard suffisamment pertinent pour ce sujet de recherche.

Certains projets ont également ouverts des voies intéressantes dans l'exploration de l'espace des hypothèses. D'abord le projet CAT a intégré un nouveau concept dans l'espace des hypothèses, créant une extension de l'espace et la possibilité d'interroger les bases de données sous un nouvel angle. Ensuite, le projet AROUND a cherché à explorer si la mise en place d'essais cliniques dans les hôpitaux avaient un impact sur la mortalité du cancer en France. La combinaison de ces deux termes {essais cliniques} et {mortalité} du cancer est une approche originale selon les scientifiques et suffisamment intéressante pour être traitée de façon isolée afin de proposer des indicateurs dans le cadre de la méta-épidémiologie et des études réalisées. Enfin, le projet *Locapred* a permis de montrer que les bases de données ouvertes utilisées permettaient d'interroger des facteurs de risques connus ou suspectés par la littérature et donc de pousser à investiguer dans ces directions.

Il est intéressant de voir que la valeur scientifique du projet *Locapred* est évaluée comme haute. Pourtant, d'un point de vue de l'épidémiologie du cancer, celui-ci n'offre pas de résultats pertinents. En fait, sa valeur scientifique a été montrée dans le domaine du machine learning et des techniques de data mining. Il y a extension de la fonction de valeur initialement construite : nous reviendrons sur ce point dans la dernière section.

	Science	Compatibilité données	Politique publique	Interaction acteurs potentiels	Originalité	Total
H1	xx	x	x	o	o	**27%**
H2	xx	x	x	o	o	**27%**
H3	o	o	o	o	o	**0%**
A1	xx	o	xx	x	xx	**47%**
A2	x	x	x	o	o	**20%**
A3	x	x	x	o	x	**27%**
A4	xxx	x	x	xx	o	**47%**
A5	o	o	o	o	o	**0%**
A6	o	o	o	o	o	**0%**
A7	o	o	o	o	o	**0%**
A8	o	o	o	o	o	**0%**
Outil de data visualisation (*IDEA*)	o	xxx	-	xx	xx	**47%**
Outil de nettoyage des données (*Locapred*)	xxx	xx	x	x	x	**53%**

Tableau 21. Analyse de la valeur par projet du challenge 2 Epidemium.

3. *Effet de la capitalisation sequentielle : extension de l'espace des hypotheses et de la fonction de valeur*

3.1. La reussite du challenge 2 portee en partie par la capitalisation des participants

Le principal changement en terme de gestion opéré entre le challenge 1 et le challenge 2 est un renforcement des contraintes en terme de résultats : en plus des thématiques suggérées, les organisateurs ont imposé aux participants de situer leurs activités vis-à-vis de la littérature scientifique. En contrepartie, ils ont mis à disposition des équipes un ensemble de spécialistes du domaine constitués d'épidémiologistes, d'oncologues ou d'acteurs de la santé publique. Si la valeur des résultats produits est dans certains projets meilleurs que dans le challenge 1, il serait précipité de dire que c'est uniquement grâce à l'action des organisateurs et des financeurs. En effet, plusieurs éléments nous portent à croire que ce n'est pas le cas.

Premièrement, nous avons vu que les rendus des étudiants ont été très décevants. En effet, les résultats obtenus par les projets étudiants sont souvent basés sur plusieurs cancers en même

temps, et ne permettent pas d'obtenir le niveau de validité exigé par la discipline. Bien que ces derniers avaient accès à un ensemble de spécialiste du domaine pour s'interroger sur leur approche, peu d'entre eux en ont effectivement tiré parti, et leur travail a été plutôt le résultat d'un vase clos. Or, les équipes étaient constituées uniquement d'étudiants spécialistes en analyse de données et n'avaient pas de compétences en épidémiologie ou de façon plus large dans les questions de santé. Au final, les organisateurs ont appris peu de choses sur la thématique 3 et les projets étudiants ont plutôt servi de vecteur de communication afin de diffuser l'existence et l'intérêt du programme Epidemium dans les cercles universitaires. Pourtant, le problème de la validité des hypothèses formulées avait déjà été identifié lors du premier challenge. Nous avions mis en avant que les équipes n'avaient pas suffisamment pris en compte les conditions de validité relative à la discipline. L'accessibilité à des spécialistes du domaine ne semble pas suffisante pour éliminer ce problème, et les organisateurs devraient annoncer explicitement ces conditions de validité en même temps que les thématiques.

Deuxièmement, les projets avec les résultats les plus aboutis (*IDEA* et *Locapred*) et qui ont été récompensé par Epidemium proviennent d'équipes qui étaient déjà présentes au premier challenge. Lors de nos discussions avec les membres de ces équipes projet, nous avons compris que leur expérience lors du premier challenge avait été capitale pour eux dans le déroulement de leur deuxième intervention. Les choix qu'ils ont réalisés, autant dans la formulation de leur problématique que dans leur rapport aux données, ont été la conséquence de l'apprentissage qu'ils avaient eux-mêmes fait de leur expérience passée. Comme nous l'avions précisé pour le premier challenge, la capitalisation sauvage entre les deux challenges n'a pas été suffisante pour rendre compte des problématiques gestionnaires mais également de la capitalisation sur la production. Nous avons montré qu'une partie de cette capitalisation avait été portée par les équipes des participants. Ainsi, le résultat du challenge aurait été tout autre si les équipes n'avaient pas reconduit leur participation, incluant le bagage d'un apprentissage tacite.

Au final, si le deuxième challenge a eu de meilleurs résultats que le premier challenge, c'est principalement grâce aux équipes qui ont elles-mêmes capitalisé sur ce qu'elles avaient appris durant le premier challenge. Or, il est nécessaire que cet apprentissage soit géré par les organisateurs.

3.2. EXTENSION DE L'ESPACE DES HYPOTHESES

Le deuxième challenge a mis en évidence l'émergence de projets comme CAT qui construisent des hypothèses indépendamment des bases de données. Au lieu de chercher à produire une hypothèse qui colle le mieux aux données disponibles, les membres de l'équipe CAT ont réalisé le processus inverse : ils ont considéré que les données existantes étaient potentiellement suffisantes pour traiter un problème que les participants avaient identifié au préalable. Ce phénomène est caractéristique d'un fort imaginaire sur la capacité des données massives. Durant notre analyse de cas, nous avons pu constater à de multiples reprises que l'existence même de bases de données

massives dont le contenu est incertain est souvent associée à une croyance sur ce qu'il est possible de créer à partir de ces données. Cela peut mener les citoyens de la science à formuler une hypothèse : le nombre de variables à l'intérieur des bases de données massives est tellement grand qu'il est statistiquement possible que une hypothèse formulée de manière *ad hoc* puisse être vérifiée par ces données. Les participants sont alors sujets à une « appétence » qui les poussent à exploiter les données comme une ressource pour étudier un sujet.

Si cette appétence ne peut être maîtrisée dans un processus totalement ouvert, elle ne doit pas être négligée par les organisateurs car elle crée une opportunité de faire émerger des questions intéressantes que l'on peut tester à partir des bases de données existantes. Nous avons vu que, contrairement à l'approche classique qui consiste à reformuler les hypothèses de départ pour qu'elles correspondent aux bases de données, la stratégie ici consiste à trouver des variables suffisantes pour expliquer l'hypothèse préalablement formulée. Dans le cas du projet CAT, la notion de « survivance » n'existe pas dans les bases de données existantes, ainsi que dans la littérature en épidémiologie du cancer. En fait, le terme permet de faire une extension de l'espace des hypothèses en intégrant une nouvelle relation entre les familles de concept existantes.

3.3. EXTENSION DE LA FONCTION DE VALEUR

Enfin, nous avons montré que si la fonction de valeur est nécessaire pour piloter l'exploration des espaces dans les tâches déléguées à la foule, elle ne peut pas être considérée comme étant inconditionnellement figée. En effet, la fonction de valeur créée durant le premier challenge n'a pas été suffisante pour rendre compte de ce qui a pu être produit durant le challenge 2. Le projet *Locapred* a par exemple débouché sur la publication de papiers scientifiques dans le domaine du machine learning, bien éloigné du domaine de l'épidémiologie du cancer et donc de la définition que l'on avait proposé de la « valeur » scientifique. Tout se passe comme si l'exploration réalisée par l'équipe *Locapred* avait étendu les critères de valeurs qui avaient été préalablement établi par les organisateurs. Nous pouvons représenter la fonction de valeur comme un ensemble de critères $\{c_1, c_2, \dots, c_n\}$ avec n fixé (5 dans notre cas durant le challenge 1). Dans la **figure 58** que nous avons utilisée pour notre étude, les valeurs sont représentées dans un espace à deux dimensions, où les valeurs fluctuent en fonction de la position dans l'espace explorée.

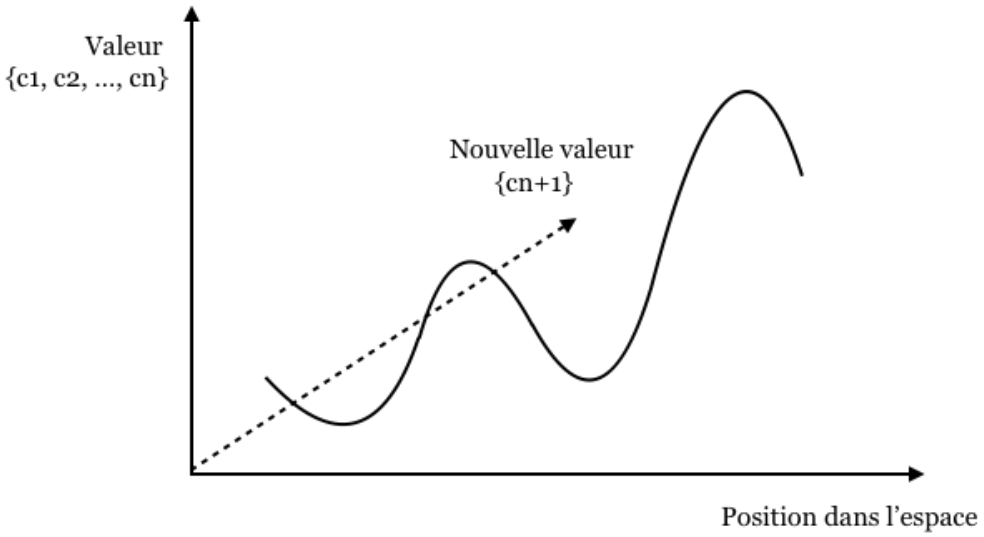

Figure 58. Nouvelle valeur $\{c_{n+1}\}$ intégrée dans les critères définis préalablement

Le projet *Locapred* permet une extension de cette fonction de valeur en intégrant la notion de « valeur scientifique dans le domaine du machine learning ». Cela représente une extension des critères de valeurs établis préalablement en ajoutant une dimension $\{c_{n+1}\}$. Sans cette extension, l'évaluation du projet serait perçue comme incompatible avec les objectifs du programme et donc écartée du processus de capitalisation séquentielle. Ainsi, en plus de gérer la capitalisation par la valeur des projets, il y a nécessité de gérer l'extension de cette valeur.