AMÉLIORATION DU PORTIQUE DE SANTÉ

PSC - Rapport final

Août 2021 - Avril 2022

MAP11

Ben Abdallah, Bettannier, Ding, Quéran, Schneider, Tran Ha

TABLE DES MATIÈRES

1 REMERCIEMENTS

Nous avons eu la chance de pouvoir nous investir dans la réalisation de ce projet grâce à l'entreprise Valeo qui a proposé ce sujet de PSC.

Nous tenons à remercier l'ensemble des membres de l'entreprise qui nous ont accompagnés tout au long de l'avancement de ce PSC. Nous pensons tout particulièrement à Mme Sylvie Cavelier, qui nous a guidé hebdomadairement dans l'avancement de notre travail au cours de nos réunions au Drahix Innovation Center. Nous remercions également M. Georges de Pelsemaeker, Tech Incubator chez Valeo, qui nous a également accompagnés dans cette étude, orientant nos recherches bibliographiques et nous fournissant les données que nous avons plus tard utilisées pour nos algorithmes.
Nos remerciements s'adressent par ailleurs à M. Philippe Goge, ingénieur et data scientist chez Valeo, pour l'aide précieuse qu'il nous a pu nous apporter dans l'implémentation de nos algorithmes, pour la compréhension du fonctionnement des modèles de machine learning employés, ainsi que pour ses conseils particulièrement efficaces. Grâce à lui, nous avons pu véritablement comprendre l'essence d'un travail quotidien dans le domaine du machine learning : loin des premières idées que nous pouvions avoir en nous lançant dans ce projet, ce travail se définit avant tout comme l'élaboration d'algorithmes adaptés à une problématique spécifique et à notre jeu de données, une part considérable du travail en question étant dédiée à l'analyse et au traitement de ces données.
Nous remercions aussi plus généralement tous les employés de Valeo dont le travail a pu nous être utile, notamment toutes ceux ayant réalisé l'acquisition des données que nous avons utilisées.

Nous remercions enfin M. Sylvain Le Corff, notre professeur référent pour avoir veillé au bon déroulement de ce PSC.

2
EXECUTIVE SUMMARY

Au cours de ce projet, nous avons choisi de travailler avec l'équipementier automobile Valeo sur un projet de portique de santé capable d'analyser à l'aide de méthodes non invasives les signes vitaux d'un patient afin de décider si celui-ci présente ou non une pathologie déterminée (par exemple la Covid-19). L'objectif était ainsi d'implémenter un algorithme de machine learning qui serait à même de prédire si le patient était malade ou non à partir des valeurs de différents paramètres biologiques fournies par les capteurs du portique santé.

Pour atteindre cet objectif d'obtention d'un algorithme prédictif fonctionnel, il a fallu sélectionner un modèle à appliquer à l'ensemble des données récupérées sur de multiples patients avant de se lancer dans l'étape incontournable d'entraînement de l'algorithme à partir de ces données. Les données en question, collectées dans plusieurs centres de mesure répartis à travers le monde, nous ont été fournies par Valeo. Une fois suffisamment de données récupérées, nous les avons analysées en profondeur afin de les nettoyer en éliminant les données inutilisables ou incomplètes qui viendraient fausser les résultats des prédictions de l'algorithme que nous cherchions à élaborer. Cette étape, qui s'est avérée beaucoup plus longue et complexe que ce que nous avions prévu initialement, a été non seulement cruciale dans l'avancement de notre projet mais également riche d'enseignements pour nous dans notre démarche scientifique.

À la suite de la mise en place de programmes de tri en mesure de sélectionner les données pertinentes pour la prédiction de l'état du patient, nous avons pu achever l'étape de nettoyage de données. Il nous a fallu ensuite nous pencher sur un choix de modèle prédictif. Nous avons débuté avec un modèle particulièrement simple : un estimateur bayésien naïf qui nous a permis d'obtenir des résultats corrects qui restaient cependant améliorables. Nous nous sommes donc efforcés d'affiner notre modèle pour améliorer son efficacité de prévision en nous tournant vers des algorithmes de forêts aléatoires avant de nous pencher sur des réseaux de neurones. Nous avons été guidés au travers de ces différentes étapes par plusieurs membres de l'équipe de Valeo avec laquelle nous avons travaillé : les connaissances qu'ils ont pu nous apporter ont été, indépendamment des résultats obtenus, particulièrement instructives, nous offrant une première introduction au monde du machine learning.

3
PRÉSENTATION GÉNÉRALE

3.1 Introduction au sujet

Afin de lutter au mieux contre la propagation de l'épidémie de Covid-19, l'équipementier automobile Valeo s'est proposé de mettre la technologie sur laquelle repose la conception de ses véhicules intelligents au service de la détection de patients infectés par le virus. En effet, l'interaction entre ces véhicules intelligents et leurs passagers requiert l'utilisation de différents capteurs biométriques permettant d'analyser en détails les signes vitaux des passagers en question (la température et la fréquence respiratoire pour ne citer qu'eux). Ces capteurs sont également couplés à des systèmes de traitement étudiant les données obtenues dans le but de déterminer la réponse du véhicule la plus adaptée.

Figure 1 – Pod utilisé pour réaliser l'acquisition des données

Ces capteurs biométriques ont par conséquent été transposés dans le cadre d'un portique de santé, capable d'effectuer un diagnostic complet des patients et de leurs signes vitaux. Dans l'urgence de la pandémie de Covid-19, ces portiques pourraient permettre de désengorger les hôpitaux en testant les patients de manière automatisée, fiable (à 95%) et avant tout rapide. Cependant, l'usage de ces portiques a le potentiel de dépasser largement celui de solution d'urgence pour venir endiguer les ravages de la pandémie : ils pourraient en effet, à l'avenir, être mis au service de la détection d'autres pathologies. L'analyse des données collectées par les capteurs doit permettre au dispositif de se prononcer quant à la présence ou l'absence de pathologies chez le patient. Nous nous sommes donc proposés de développer un portique de santé en mesure de détecter une autre pathologie que la Covid-19.

Nous avons cherché dans un premier temps à définir les grandeurs biologiques que nous pouvions envisager de mesurer avec l'aide des capteurs biométriques présents dans le portique santé. Nous nous sommes notamment intéressés à l'oxymétrie, l'alcoolémie, la glycémie, la température, la pression artérielle et au rythme cardiaque. De même, nous nous sommes renseignés sur les moyens utilisés pour étudier ces paramètres chez les patients, à savoir les caméras thermiques, les microphones ou encore les radars à ondes courtes modulées. Nous avons alors cherché à nous faire une idée de l'état de l'art concernant ces différents paramètres, à nous renseigner sur les modèles théoriques déjà disponibles pour étudier les caractéristiques biologiques des patients mais aussi (et surtout) de la faisabilité expérimentale de ces études. Nous sommes donc passés par la case de recherche bibliographique, en épluchant les propositions de la littérature scientifique à ce sujet afin de mieux nous approprier ces thématiques connexes à notre projet.

3.2 Présentation de la littérature et analyse de la concurrence

En lien avec Valeo, nous avons débuté notre PSC par une analyse concurrentielle de produits préexistants offrant certaines fonctionnalités que le portique santé doit inclure. Cela nous a permis de récolter quelques informations sur différents produits commercialisés par environ 80 entreprises. La plupart sont des startups bien que l'on retrouve aussi dans le lot de grandes entreprises comme Google, Facebook, Amazon ou encore Apple.

Nos recherches ont ainsi indiqué que plusieurs méthodes non invasives avient déjà été développées afin de détecter différents signes vitaux d'une personne. Citons par exemple : Wifi BodyScale de l'entreprise Withings qui mesure la fréquence cardiaque, la température et la pression artérielle ; Smart Seat Cushion de l'entreprise VISSEIRO qui mesure la fréquence cardiaque et sa variabilité, la fréquence de la respiration, l'actigramme et le niveau du stress ou encore l'Apple Watch 6 d'Apple qui permet de mesurer le taux d'oxygène dans le sang.

Dans un même temps, nous avons mené un travail de recherche en nous plongeant dans la littérature scientifique dédiée à la mesure des signes vitaux et plus précisément à la mesure du SpO2 (saturation pulsée en oxygène). La saturation en oxygène correspond au taux d'oxygène contenu dans les globules rouges après leur passage dans les poumons. Cela permet d'évaluer les fonctions respiratoires d'une personne. Aujourd'hui, pour mesurer le SpO2, on utilise un saturomètre, un petit appareil que l'on place au bout de son doigt. Bien que cette méthode soit peu invasive, les ingénieurs de Valeo qui nous accompagnent sur ce projet nous ont confié leur objectif d'obtenir une méthode de mesure du SpO2 qui opérerait en l'absence totale de contact.

Pendant notre période de recherche au sujet du calcul et de la mesure du SpO2, nous avons pu nous appuyer sur la documentation interne sur le portique santé, fournie par Valeo, mais également sur des articles universitaires. Grâce à ce travail, nous avons pu remarquer que l'oxymétrie présentait plusieurs intérêts au nombre desquels la détermination de la sévérité d'une

maladie ou d'une détresse respiratoire (hypoxémie), le dépistage de la Covid-19 (il y a en effet une forte corrélation entre une mesure de SpO2 inférieure à 95% et un test positif à la Covid) ou la prédiction de l'évolution d'un patient infecté (forme grave ou pas) afin de déployer un traitement adapté.

Nous avons également rencontré à travers de ces articles plusieurs méthodes utilisées pour la mesure de la saturation en oxygène avec une caméra RGB (Red, Green, Blue) qui nous ont été utiles, notamment la régression linéaire, les réseaux de neurones et les forêts aléatoires. L'étude de ces différents travaux nous a permis d'être alertés à propos de certaines difficultés auxquelles nous pourrions faire face comme l'impact de la couleur ou de l'épaisseur de la peau, l'éclairage qui altère la couleur et la précision des images exploitées par les algorithmes ou encore l'environnement de mesure (température ou modification du rythme cardiaque) : autant de caractéristiques à même de modifier le volume des vaisseaux sanguins et ainsi de changer leurs propriétés d'absorbance.

Ces multiples éclairages nous ont permis de cadrer notre projet autour d'une problématique centrale : s'appuyer sur des algorithmes de machine learning afin de déterminer le SpO2 de patients à distance, grâce à des images RGB et à des données brutes qui se trouveraient dans les différentes bases de données que Valeo nous proposait.

4
ANALYSE DES DONNÉES

Afin de mettre en place un algorithme de prédiction du SpO2 chez les patients, nous nous sommes appuyés sur un ensemble de résultats de mesures auxquels Valeo nous avait accordé l'accès. L'entreprise dispose de plusieurs sites de collecte de données disséminés à travers le monde et qui ont été mis à contribution afin d'obtenir notre set de travail. Parmi ces différents sites de collecte de données, le plus important est situé au Maroc. Les autres sont situés en Inde, à Malaga, au CHU Henri Mondor (qui pourrait présenter l'avantage de nous fournir des mesures réalisées sur des patients en bien moins bonne santé que dans la majorité des autres centres de mesure) ainsi qu'au Drahix Innovation Center (où nous nous sommes prêtés au jeu et avons effectué des acquisitions sur nous-mêmes). Les résultats des mesures effectuées sur ces sites nous permettaient ainsi, début mars 2022, de disposer d'un jeu de données regroupant les informations de 8684 personnes (voir figure 2).

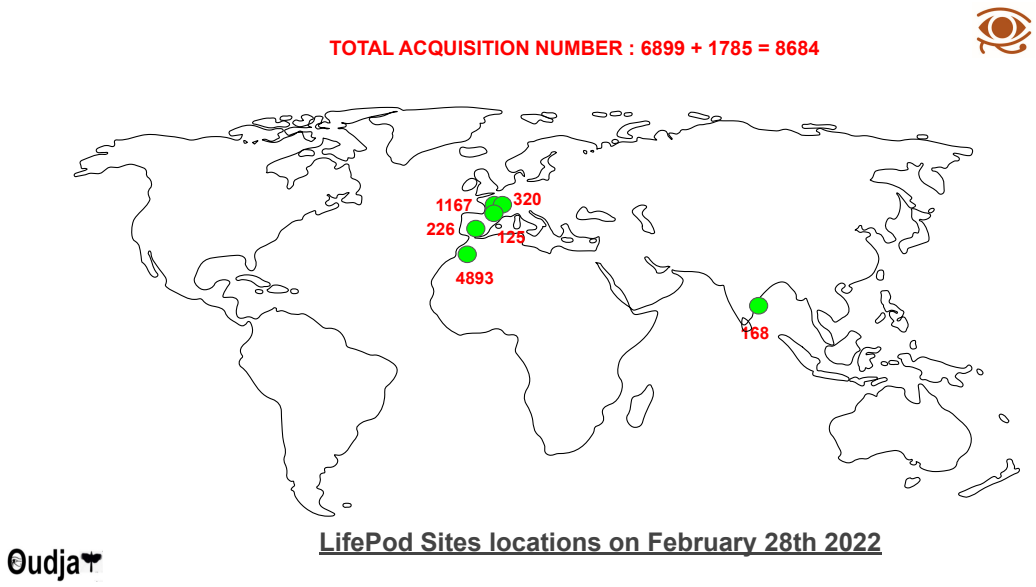


Figure 2 – Localisation des sites d'acquisition de données

Pour obtenir ces données, l'entreprise avec laquelle nous avons travaillé utilise une caméra RGB (Red, Green, Blue). La technologie employée par cette caméra s'appuie sur le principe de synthèse additive des couleurs, principe selon lequel l'ensemble des couleurs du spectre visible peut s'obtenir en combinant les trois couleurs primaires que sont le rouge, le vert et le bleu. Ce modèle permet ainsi de coder les informations de couleurs détectées par ces caméras sur les patients. La caméra concentrait les mesures sur des landmarks qui sont des points de repère répartis sur le visage.
Pour chacun de ces points de repère, le capteur fait l'acquisition de différentes données au nombre desquelles on retrouve aussi bien un codage RGB indiquant la couleur du point que des

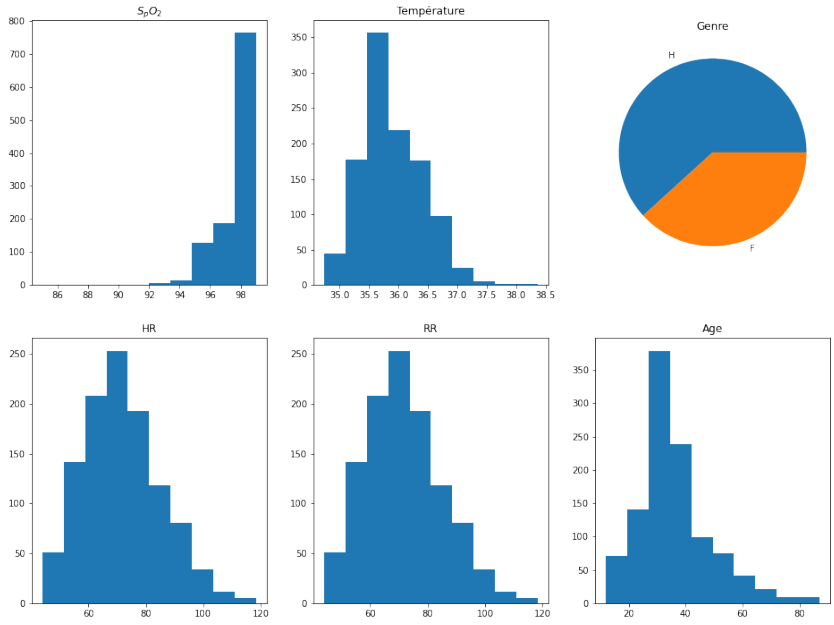
coordonnées permettant de situer le point sur le visage ou encore diverses informations relatives à la température (températures maximales, températures moyennes et températures brutes par exemple).

4.1 Présentation du dataset

Nous avons reçu au total 22.7 GB de données de la part de Valeo. Ces données qui nous ont été fournies étaient réparties en un ensemble de dossiers, chaque dossier correspondant à un patient différent. À chaque patient est associé un groupe de mesures de ses signes vitaux, mesures qui sont effectuées dans un premier temps par le portique santé puis qui sont complétées par un opérateur médical préposé à cette tâche. Ces données mesurées se répartissent en deux types de fichiers d'extension TDMS : d'une part des fichiers RawData accompagnés d'un index et d'autre part des fichiers FaceMask, eux aussi accompagnés d'un index. Il est également intéressant de remarquer qu'en règle générale un fichier FaceMask est beaucoup plus volumineux qu'un fichier RawData : classiquement environ 75 kB pour ce dernier contre 1.4 MB pour le premier.

SpO2	Température	HR	RR	Genre	Age
1102	1102	1097	1097	771	1086

(a) Nombre de données exploitables



(b) Distribution

Figure 3 – Étude statistique des premières données (hors masque RGB)

Nous pouvons observer sur la figure ci-dessus la distribution de nos données RawData.
La première chose que nous remarquons est le faible nombre de données exploitables : à peine plus de 1000 données exploitables pour chaque catégorie (à l'exception du genre).
Nous voyons également très clairement une forte concentration des données : le SpO2 est concentré à 98 par exemple et les personnes sont âgées principalement entre 25 et 35 ans pour ne citer que ces deux catégories.

Ainsi, le dataset présente un biais de classe très important : l'immense majorité des personnes dont les résultats forment notre ensemble de données sont en bonne santé.
Par conséquent, les signes vitaux sont somme toute relativement semblables tout au long du dataset.
En outre, du fait des environnements au sein desquels les acquisitions des données ont été réalisées, même les données qui ne correspondent pas à des signes vitaux sont relativement homogènes (genre, âge).

Quant aux fichiers FaceMask, ils se présentent sous la forme de tenseurs de données et sont structurés en plusieurs colonnes dédiées respectivement aux :
- indexations des mesures ;
- nombres de landmarks du patient pour lesquels des mesures ont été effectuées ;
- codes RGB du landmark ;
- coordonnées de ce point de repère sur le visage du patient, définies à partir d'une abscisse X et d'une ordonnée Y ;
- diverses mesures de températures (T°disp, T°moyenne(max), T°max(max), T°brutes).

Nous avons dû dans un premier temps convertir ces fichiers TDMS en fichiers CSV afin de simplifier leur lecture et leur traitement. Chaque visage regroupait jusqu'à 469 landmarks (lorsque l'acquisition était bien accomplie). Ils permettent de couvrir l'ensemble du visage. Notons toutefois qu'ils sont plus nombreux à certains endroits qu'à d'autres. En effet, la densité des landmarks au niveau du contour des yeux, du nez et de la bouche est bien plus importante que dans le reste du visage.
Cela est lié au fait qu'il s'agisse de zones stratégiques où les variations des données à récupérer sont bien plus marquées. Plus de capteurs sont donc requis à ces endroits pour mieux rendre compte de cette évolution.

De plus, comme l'on sait à l'avance que certaines des données sont incomplètes voire inutilisables, il nous faut pouvoir les détecter rapidement. Nous avons pour cela implémenté un algorithme de machine learning qui détecte si un masque correspond bien à un visage utilisable ou pas. Nous reparlerons de cet algorithme dans la section qui lui est dédiée. Mais, le fait est que, pour pouvoir reconnaître s'il s'agit bien d'un visage, les landmarks doivent occuper des endroits stratégiques. Ainsi la forte concentration de ces points dans des zones importantes telles que le contour des yeux, le nez et la bouche est un facteur primordial à prendre en compte pour notre travail sur ces données.

Toutefois, en raison du positionnement variable des individus devant la caméra ainsi que

d'éventuels mouvements du patient pendant le processus d'acquisition, toutes les zones du visage ne disposent pas de la même accessibilité pour le capteur. Il en découle que certaines zones sont potentiellement mal positionnées, voire masquées. Cette superposition des landmarks est donc à l'origine de résultats brouillés, difficilement utilisables pour notre étude. Ce problème est à l'origine de la majorité des données inexploitables.

Une fois cette première étape de visualisation des données effectuée, il nous restait encore à les nettoyer pour pouvoir entraîner nos modèles avec.

4.2 Le nettoyage des données

4.2.1 • Les données RawData

Ces données nous ayant été fournies en format brut : il nous a fallu tout d'abord les nettoyer. Une étude plus approfondie de ces dernières a été nécessaire afin de mieux comprendre leur structure, ce qu'elles représentaient ainsi que la cohérence et la pertinence des informations qu'elles apportaient à notre étude.

De ce fait, dans un premier temps, nous avons éliminé à la fois les résultats lacunaires et ceux dont la valeur était manifestement absurde. L'inconvénient d'une telle pratique est qu'il ne nous restait qu'un nombre fortement restreint de données : 754 sur les 8684 que nous avions reçues directement de Valeo, soit seulement 8.7% des données totales.

Afin d'avoir davantage de données utilisables, nous avons essayé de compléter les jeux de données lacunaires. Nous avons commencé par essayer de remplir par nous-même de manière réaliste les données manquantes en procédant par symétrie et en recoupant avec les jeux de données complets pour y trouver des similitudes.

En remarquant que notre méthode s'apparentait à de l'apprentissage sur un jeu de données, nous nous sommes penchés sur des algorithmes de machine learning capables d'exercer cette tâche. Nous avons utilisé le modèle des k plus proches voisins qui se prête très bien à ce type d'exercice. Il s'agit d'un algorithme d'apprentissage supervisé qui, à partir de données d'entraînement (dans ce cas les données complètes), va estimer les résultats des données manquantes. L'algorithme des k plus proches voisins fonctionne de la manière suivante :
- Le nombre k est fixé. L'algorithme considère les k voisins les plus proches (pour une certaine distance) de l'entrée considérée. Pour faire cela, nous avons considéré la distance euclidienne classique pondérée par des coefficients de corrélation afin de donner plus d'importance aux paramètres qui sont fortement corrélés avec la variable recherchée et ainsi d'avoir un résultat plus en accord avec les observations de notre jeu de données complet.
- Une fois ces k plus proches voisins détectés, on attribue la valeur manquante par vote. Ici cela correspond à une simple moyenne des valeurs prises par les k plus proches voisins.

Afin de choisir un bon k, nous avons procédé par cross validation avec les données complètes, c'est-à-dire que nous avons séparé nos données aléatoirement en plusieurs groupes de même taille puis nous avons fait tourner l'algorithme des k voisins en utilisant des groupes différents à chaque fois comme groupe test ; les autres groupes servant à l'apprentissage. Nous avons alors répété l'algorithme des k voisins en prenant différentes valeurs de k et en choisissant celle qui minimise les erreurs sur notre ensemble de tests.
En effectuant cela, nous avons vu que le paramètre $k = 5$ était le mieux adapté. Nous avons donc choisi de réaliser notre travail avec cette valeur de k.

Finalement, grâce à cet algorithme, nous avons réussi à augmenter notre nombre de données exploitables à 1225. Ainsi, sur les 8684 patients qui se sont portés volontaires pour la campagne de tests de Valeo, nous avons obtenu (seulement) 14.1% de données exploitables : nous sommes conscients que cela représente malheureusement une quantité de données somme toute assez faible pour entraîner des modèles de machine learning.

En outre, en testant une première fois nos algorithmes sur ces données, nous nous sommes rendu compte qu'il fallait recalibrer la température en fonction du pays où s'est effectuée la mesure. En effet, la position géographique influe fortement sur la température de la pièce au sein de laquelle se déroulent les acquisitions de données. Lorsque le Pod effectue ses mesures de température, nous avons remarqué qu'il était ainsi biaisé par la température de la salle.

De plus, lorsque nous nous sommes nous-mêmes prêtés au jeu d'acquisition des données, nous avons remarqué que l'appareil chauffait lors de son utilisation. Cela peut entraîner une différence de température entre le moment de la première acquisition de la journée et celle de la dernière. Malheureusement, dans les données qui nous ont été transmises, les horaires des acquisitions ne sont pas présents. Nous ne pouvions donc pas travailler sur nos données pour voir de quelles options nous disposions pour régler ce problème.

Afin de recalibrer la température, nous avons décidé d'harmoniser les données venant des différents pays dans le but d'avoir une moyenne et un écart-type similaires pour chaque centre d'acquisition à l'exception du CHU Henri Mondor. Dans ce dernier centre de tests, il nous semble normal que les données soient sensiblement différentes étant donné l'environnement particulier au sein duquel elles ont été acquises.

4.2.2 • Les données FaceMask

Nous avons également travaillé sur le nettoyage des données FaceMask. Parmi les problèmes auxquels nous avons été confrontés avec ce jeu de données figure la superposition des landmarks liée à des positionnements variables des individus devant la caméra. Ce problème est à l'origine de la majorité des données inexploitables. Afin de conserver dans notre set uniquement les données utilisables, nous avons réalisé un algorithme qui apprend à détecter si les données sont utilisables, c'est-à-dire si le visage est reconnaissable et si les données en entrée ne sont pas absurdes.

En plus de la faible quantité de données exploitables s§ont venus s'ajouter les obstacles des biais de classement, d'erreur systématique de mesure de l'exposition ou de la maladie. En effet, comme nous l'avons dit pour les RawData, nous avons remarqué que les individus permettant de construire le dataset constituaient une population principalement composée d'hommes jeunes, en bonne voire très bonne santé. De ce fait, il n'est donc pas surprenant de remarquer que la grande majorité des valeurs de SpO2 sont supérieures à 98, ce qui réduit considérablement le nombre de cas pertinents pour entraîner notre algorithme de prédiction.

Nous avons aussi fait face à la difficulté de l'éclairage. Ce dernier conditionne en effet de manière systématique les résultats d'une mesure de couleur effectuée sur un système donné : la répartition des composantes rouge, bleue et verte est amenée à varier en fonction de l'éclairage. On pourrait penser que ce paramètre est avant tout dépendant du moment de la journée auquel les tests sont effectués : dans ce cas, synchroniser la batterie de mesures viendrait assez simplement remédier à ce problème. Cependant, nous avons pu constater que les centres de mesure sur lesquels s'appuyait Valeo se trouvaient à des endroits très différents du globe : ils sont donc soumis à des conditions de luminosité très diverses. De plus, le matériel d'éclairage utilisé dans les différents centres d'acquisition n'est pas identique. Pour faire face à ces problèmes de luminosité, il nous a été conseillé de travailler avec des ratios de couleurs et notamment avec le système de représentation YCgCr que nous expliquerons dans le paragraphe qui lui est consacré, et qui est particulièrement adapté à l'étude du SpO2.

Le temps investi dans l'identification de ces différentes problématiques a été riche d'enseignements : cela a été l'occasion de mieux cerner les multiples étapes et mécanismes s'enchaînant au cours d'un projet d'intelligence artificielle. Cette démarche d'analyse des données que nous avions à disposition a clairement fait ressortir toute l'importance que peut revêtir la collecte, la préparation et le traitement des données pour nos objectifs. Ce stade est tout simplement primordial afin d'obtenir la matière première nécessaire à l'entraînement de nos algorithmes. Cependant nous avons dû nous rendre à l'évidence : cette collecte des données, loin d'être une simple formalité, peut prendre un certain temps avec des données parfois limitées qui n'arriveront que tard dans l'avancement du projet. Une coordination minutieuse entre les équipes d'acquisition des données et celles dévouées à l'élaboration des algorithmes de machine learning est donc essentielle afin que les algorithmes puissent être développés en amont de la réception des données : ceci devrait permettre de les confronter immédiatement à ces dernières et de commencer dès leur réception la phase d'entraînement.

Nos accompagnateurs chez Valeo ont alors été pour nous d'une aide précieuse. Ils furent de véritables intermédiaires entre notre groupe de travail et les équipes de Valeo préposées aux mesures dans les différents centres de collecte des données cités plus haut. Nous avons pu, grâce à eux, nous forger une première idée de ce à quoi les données sur lesquelles nous allions travailler ressemblaient avant qu'ils ne nous les transmettent une fois qu'elles leur parvenaient.

De plus, l'importance du réseau mondial de Valeo nous a permis de travailler et d'être aidés

dans notre travail d'analyse et de nettoyage des données par des équipes spécialisées. Ainsi, nous avons pu travailler avec une équipe située à Chennai en Inde en leur faisant parvenir des fichiers FaceMask incomplets ou incohérents. Nous avons également partagé avec eux notre algorithme de reconnaissance des données FaceMask utilisables.

Toutefois, il faut reconnaître que, bien que nous permettant d'avoir un set de données bien plus diversifié, la multiplicité des centres d'acquisition a également compliqué notre démarche. En effet, en raison des différences d'avancement dans les mesures entre les centres de collecte, ces données arrivaient au fur et à mesure nous contraignant à élaborer des algorithmes avec une quantité insuffisante de matière première quitte à devoir les ajuster une fois des données supplémentaires récupérées.

De plus, les résultats des acquisitions n'étant pas présentés sous la même forme pour chaque centre d'acquisition, nous devions réaliser un travail important à chaque nouvelle réception de données pour les trier et les réarranger.

5
MODÈLES ET CONCEPTS

Comme mentionné précédemment, nous avons dû élaborer des modèles d'algorithmes de machine learning en ayant peu voire très peu de données pour les entraîner. Puis, après les avoir reçues, analysées et nettoyées, nous avons pu entraîner nos algorithmes avec, ce qui nous a permis de déterminer leur efficacité afin d'éventuellement effectuer les changements qui s'imposaient.

Nous avons orienté notre travail sur les algorithmes autour de trois grands axes sur les données dont nous disposions. Le premier était de réaliser un modèle permettant de reconnaître les visages utilisables, modèle qui nous a été particulièrement utile pour trier ce type de données. Puis nous avons travaillé avec des modèles sur les deux types de données dont nous disposions : les fichiers RawData et FaceMask.

5.1 Modèles sur les RawData

Nous avons été accompagnés dans notre travail de modélisation et de mise en place d'algorithmes par un manager en data science chez Valeo. Il nous a expliqué qu'il était préférable que nous travaillions sur plusieurs modèles de difficultés croissantes afin de voir les éléments supplémentaires d'amélioration apportés à chaque étape. De plus, cette méthode de travail nous permettrait, en cas d'échec avec une méthode relativement complexe, d'identifier rapidement où l'erreur s'était produite et ainsi la corriger plus aisément.

De ce fait, nous avons dans un premier temps considéré des algorithmes plutôt simples comme la régression linéaire ou l'approche bayésienne naïve avant de nous pencher sur un algorithme plus complet et satisfaisant : l'algorithme Random Forests.

5.1.1 • Régression linéaire

Tout d'abord, nous nous sommes appuyés sur le modèle de régression linéaire. Les avantages de ce modèle étaient tous trouvés : il s'agit d'un modèle simple, facile à comprendre et rapide à entraîner.

À partir de la variable cible (ici le SpO2), le modèle utilise des variables prédictives (ici celles stockées dans les fichiers RawData) afin d'effectuer une prédiction de la cible.
Comme nous étions ici en présence de plusieurs variables explicatives (X_i) (genre, âge, fréquence cardiaque, fréquence respiratoire, température...), nous avons réalisé une régression linéaire multiple, c'est-à-dire que nous avons essayé d'exprimer le SpO2 comme combinaison linéaires des

variables explicatives.

Le résultat est donc de la forme :

$$SpO2 = \sum \lambda_i X_i + \mu + \varepsilon$$

où ε est une variable aléatoire qui représente l'erreur.

Puis, afin d'améliorer notre modèle, nous nous sommes penchés sur l'importance des features dans la régression linéaire. Le but était alors de regarder quelles features jouaient un rôle conséquent dans l'estimation de notre variable cible : le SpO2.

Genres non fusionnés

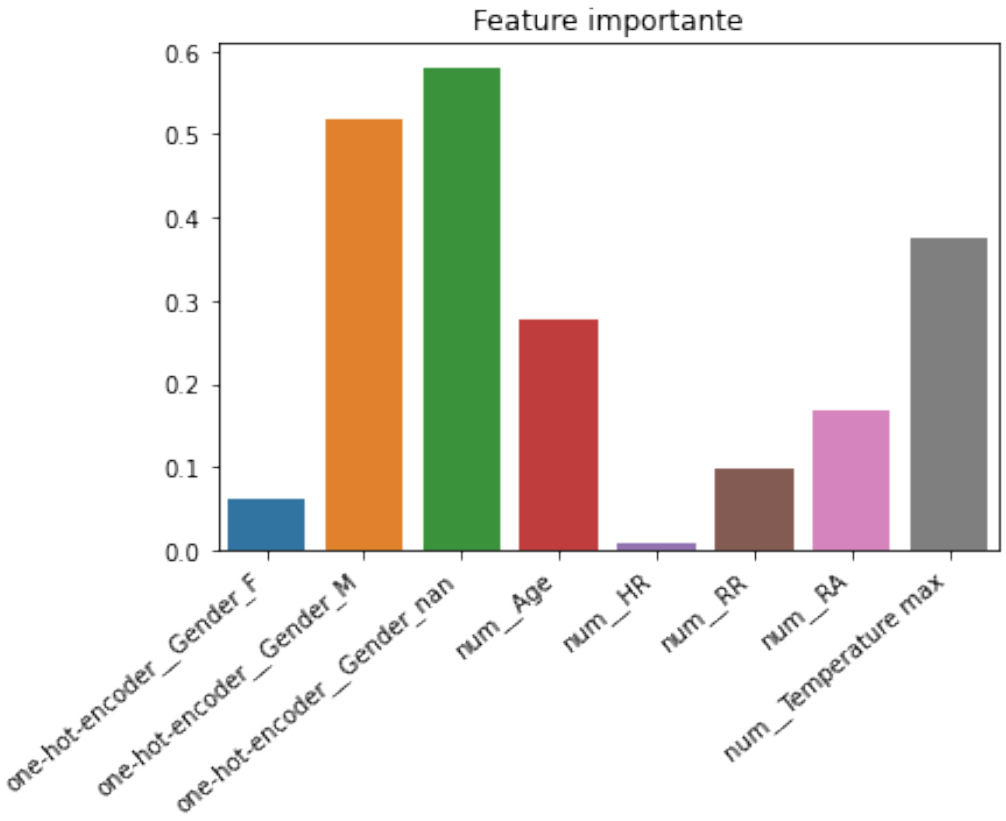


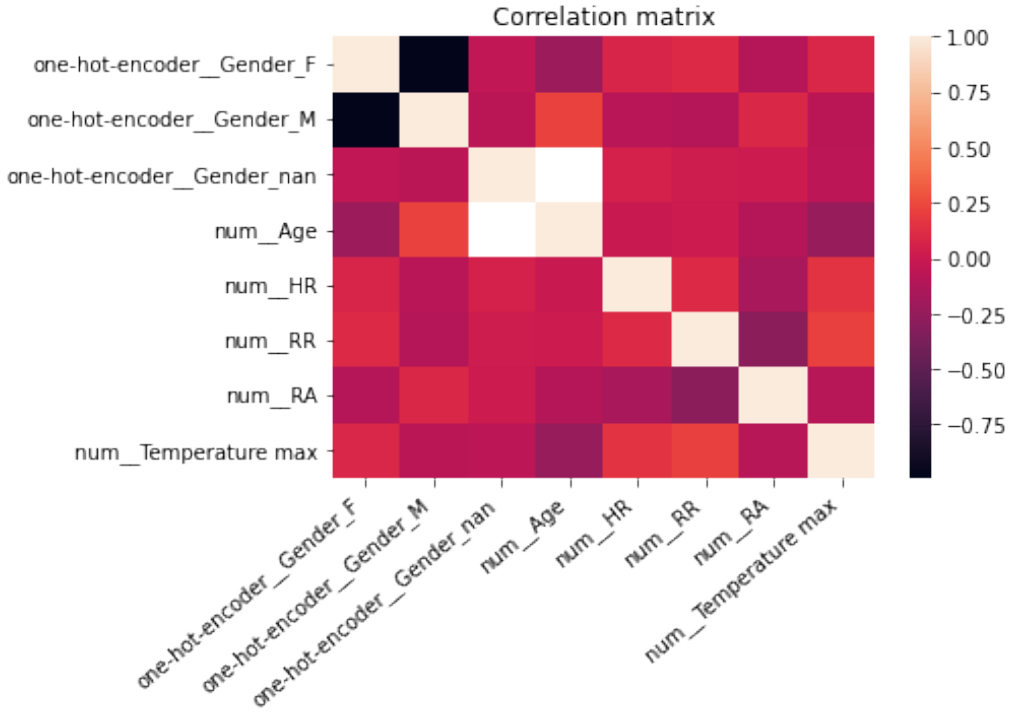
Figure 4 – Importance des variables

Nous remarquons ainsi que les features prépondérantes sont : le genre, l'âge et la température. Ces paramètres auront donc un rôle primordial à jouer dans nos algorithmes d'estimation du SpO2.
De plus, nous pensions initialement observer que certaines features pouvaient être délaissées, ou en tout cas mises de côté par rapport aux features que nous considérions comme importantes, car ayant un coefficient de régression linéaire potentiellement négligeable par rapport

aux autres. Toutefois, nous avons remarqué que cela n'est pas clairement le cas et qu'à ce point de notre travail nous ne pouvions pas nous permettre de négliger certaines données.

Toujours en poursuivant notre objectif de classification des features selon leur importance dans la détermination du SpO2, nous nous sommes également penchés sur les corrélations qu'il pouvait y avoir entre elles. Nous avons ainsi tracé la matrice de corrélation de ces entrées qui est la suivante :

Genres non fusionnés



Les données fournies initialement comportent trois entrées différentes pour le genre : homme, femme et non renseigné. Nous avons trivialement remarqué que les genres masculin et féminin sont très fortement corrélés négativement. De plus, il n'est pas intéressant de considérer des genres non renseignés. Ainsi, il suffit de considérer seulement une seule feature donnant le genre. Nous avons donc fusionné les trois genres (masculin, féminin et non renseigné) en une seule colonne. En revanche, nous remarquons qu'il n'y a pas d'autres catégories de données fortement corrélées.

Voici le graphique représentant l'importance des features et la matrice de corrélation avec la nouvelle colonnes "Gender".

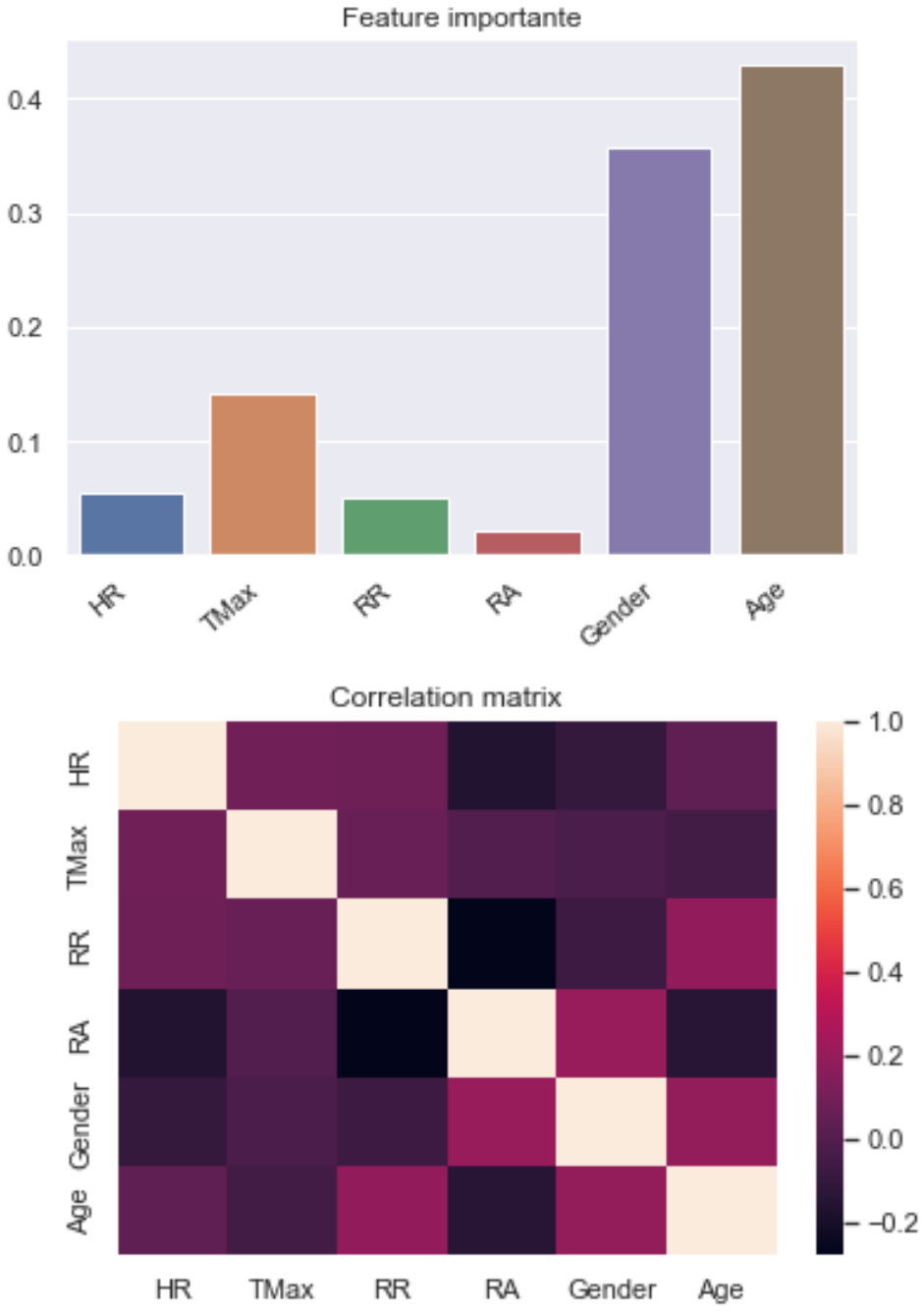


Figure 5 – Genres fusionnés

Nous remarquons que les variables explicatives qui jouent le plus grand rôle sont toujours l'âge, le genre et la température même si leurs importances respectives ont été modifiées.

5.1.2 • Bayes naïf

Suite à notre modèle de régression linéaire décrit précédemment, nous avons décidé de travailler sur des algorithmes de classification. En effet, cette approche est plus simple pour notre modèle de prédiction de SpO2. La précision extrême n'étant pas recherchée, l'objectif est de pouvoir classer la sortie de SpO2 dans un intervalle (à 2 ou 3 points de pourcentage près) ou de l'approximer à l'entier le plus proche.

Du fait de notre travail préliminaire sur l'analyse et le nettoyage des données, nous avons pu tester le modèle de classification naïve bayésienne, un modèle simple prédisant le taux de SpO2 à partir des données.

Le classificateur bayésien naïf est un classificateur linéaire, c'est-à-dire que le résultat en sortie est obtenu par combinaison linéaire des données reçues en entrée. Le classificateur bayésien naïf repose sur le modèle de maximum de vraisemblance. L'algorithme renvoie donc la réponse qui est statistiquement la plus probable parmi les classes de possibilités proposées (ici les entiers inférieurs à 100). D'où le nom de classificateur : le modèle trie les sorties dans des classes prédéfinies, ce qui est totalement différent du modèle de régression linéaire étudié précédemment.

On trouve ci-dessous une modélisation mathématique du classificateur bayésien naïf que nous avons implémenté. Dans notre cas, Y correspond au taux de SpO2 (arrondi à l'entier pour être dans une situation de classification et non de régression), et les X_i sont nos variables explicatives (à savoir le genre, l'âge, la température, le rythme cardiaque, le rythme respiratoire et l'amplitude respiratoire).

$$\text{classificateur}(x_1, \ldots, x_n) = \underset{y}{\text{argmax}}\ P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n)$$

On décompose ensuite à l'aide de la formule de Bayes :

$$P(Y = y \mid X_1, \ldots, X_n) = \frac{P(Y = y) \cdot P(X_1, \ldots, X_n \mid Y = y)}{P(X_1, \ldots, X_n)}$$

En supposant l'indépendance des variables aléatoires $X_1, \ldots, X_n$, on obtient :

$$P(X_1, \ldots, X_n \mid Y = y) = \prod_{i=1}^{n} P(X_i \mid Y = y)$$

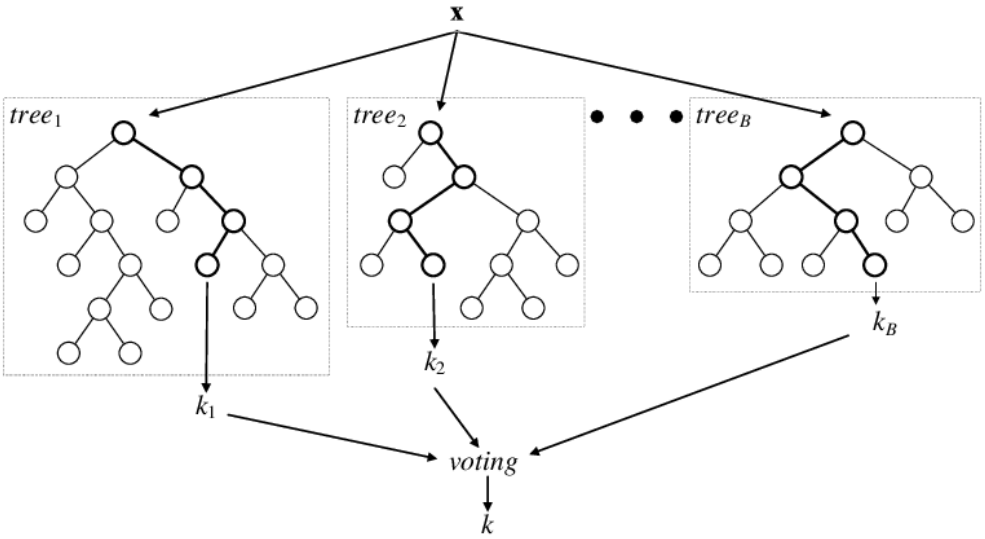
D'où finalement :

$$\text{classificateur}(x_1, \ldots, x_n) = \underset{y}{\text{argmax}}\ P(Y = y) \cdot \prod_{i=1}^{n} P(X_i = x_i \mid Y = y)$$

En autorisant une marge d'erreur de 2%, nous obtenons un taux de précision de notre classificateur bayésien naïf de 86%. À première vue, ce résultat pourrait paraître satisfaisant. Toutefois, après une réflexion plus approfondie sur notre modèle et nos données, nous pouvons

établir l'insuffisance de ce modèle. Ce taux n'a en effet pas de sens au vu des importants biais de classe intervenant ici dont nous avons parlé dans la section de présentation de données : genre, âge, le fait que quasiment toutes les personnes sont en bonne santé, etc.

5.1.3 • Random Forests



Source : https://www.researchgate.net/figure/A-general-architecture-of-a-random-forest_fig1_281175132

Figure 6 – Architecture d'une forêt aléatoire

Notre accompagnateur en data science sur ce projet nous a ensuite présenté des modèles d'algorithmes de machine learning plus sophistiqués parmi lesquels figure notamment l'algorithme Random Forests.

Pour présenter cette méthode, il convient d'évoquer dans un premier temps les arbres de décision. Les modèles d'arbres de décision, comme le modèle d'algorithme Random Forests, correspondent à des méthodes non paramétriques.

Les méthodes non paramétriques se réfèrent à un ensemble d'algorithmes qui ne font aucune supposition par rapport à la fonction cible à estimer. Ainsi, la fonction estimée ne prend pas une forme prédéterminée comme cela avait pu être le cas avec notre algorithme de régression linéaire par exemple. Le résultat est donc construit uniquement sur les informations que l'algorithme trouvera dans les données. Notons que cela est intellectuellement plus satisfaisant qu'une méthode paramétrique. En effet, nous ne savons pas à l'avance quelle forme est supposée prendre la courbe représentative du SpO2 ; l'hypothèse de la régression linéaire n'étant qu'une hypothèse simplificatrice qui avait pour but de tester un modèle.

Cependant, il est important de noter que les modèles non paramétriques présentent également des inconvénients. Le principal étant que de tels modèles nécessitent un plus grand nombre de données que les modèles paramétriques. En effet, la structure du résultat n'étant pas connue à l'avance, l'algorithme doit d'abord estimer cette structure en plus d'estimer les paramètres. Or, comme nous l'avons vu dans la présentation des données, le nombre de données exploitables que nous avons recueillies est somme toute assez faible.

Un arbre de décision est une séquence de simples règles de décisions. L'algorithme descend dans l'arbre au fur et à mesure et exécute les règles de décision. Le contrôle de la profondeur de l'arbre nous permet de trouver un équilibre entre précision et temps d'exécution.

Un algorithme d'arbre de décision est donc clair à comprendre à première vue (voir figure 6).

Le modèle des Random Forests se fonde sur les arbres de décision mais, comme son nom l'indique, plusieurs arbres de décision seront mis en jeu.
On prend d'abord plusieurs ensembles de valeurs données qui nous servent à échantillonner notre modèle. Pour chaque ensemble, on va entraîner un arbre de décision avec les données de cet ensemble. De cette manière, on obtient un ensemble d'arbres qui constituent notre forêt. Puis, une fois notre forêt constituée, l'algorithme effectue, afin de prédire la cible, un vote majoritaire entre les prédictions de chaque arbre de la forêt pour nous renvoyer son estimation. Dans un tel algorithme tous les arbres ont un rôle symétrique : il n'y a pas d'arbre pour cacher la forêt. Ce modèle permet donc de regrouper les résultats des arbres afin d'offrir une estimation plus fiable. La profondeur des arbres et leur nombre sont les paramètres qui permettent d'effectuer un arbitrage précision entre complexité.

La structure de forêt est toutefois plus difficile à représenter que celle d'arbre. Ainsi, bien que le principe de l'algorithme soit assez simple à comprendre, la représentation de la forêt et la manière de réaliser des estimations peuvent être bien plus compliquées.

5.2 Modèle FaceMask

Après ce travail sur les RawData, nous avons travaillé sur les échantillons de FaceMask utilisables, échantillons que nous avons détectés grâce à un algorithme que nous avons implémenté et que nous présenterons plus loin (car reprenant les notions des modèles que nous avons utilisés pour prédire le SpO2 à partir des fichiers FaceMask). Notre objectif est donc d'utiliser ces données des visages complets pour établir un modèle nous permettant d'estimer le SpO2.

Étant donné que les données d'un FaceMask sont plus complexes que les RawData, une utilisation des algorithmes sur lesquels nous nous sommes appuyés précédemment pour les RawData, quand bien même nous les adaptions à notre nouveau type de données, ne serait pas satisfaisante. De ce fait, une méthode tenant compte de cette complexité des données nous a semblé nécessaire. Nous nous sommes alors penchés sur des modèles de réseaux de neurones qui

nous paraissaient mieux correspondre au type des données en entrée, plus difficiles à manipuler que des tableaux de chiffres. De plus, cette difficulté ne fait que renforcer le rôle fondamental du traval préliminaire d'analyse des données, notamment la détection des principales zones d'intérêt des landmarks (coins des lèvres, yeux etc.) qui fut la première recommandation de notre encadrant chez Valeo.

Au regard des données à notre disposition, il nous semble également intéressant d'essayer de classifier les résultats en sortie entre la classe "le SpO2 est supérieur à 97" et "le SpO2 est inférieur ou égal à 97" qui représentent deux grandes catégories qu'il serait pertinent d'étudier. La première regroupe environ 60% des données et la seconde les 40% restantes. En effet, le manque de données disponibles nous contraint à réaliser cette séparation, certes peu précise, mais utile tout de même.

Notre accompagnateur en data science de chez Valeo nous a également conseillé de travailler sur un système de représentation YCgCr plutôt qu'un système de représentation RGB. Dans cette représentation, Y est le signal de luminance tandis que Cg et Cr correspondent à des données sur la chrominance : Cg correspond au complémentaire du signal de luminance par la couleur verte et Cr correspond au complémentaire de ce signal de luminance par la couleur rouge.

Voici les formules permettant de passer du système RGB au système YCgCr :

$$Y = 16 + \frac{65.481 * R}{255} + \frac{128.533 * G}{255} + \frac{24.966 * B}{255}$$

$$Cg = 128 - \frac{81.085 * R}{255} + \frac{112 * G}{255} - \frac{30.915 * B}{255}$$

$$Cg = 128 + \frac{112 * R}{255} - \frac{93.786 * G}{255} - \frac{18.214 * B}{255}$$

Pour commenter et comprendre ces formules, il est nécessaire de souligner que les composantes dans la représentation RGB sont des entiers compris entre 0 et 255, d'où la division par 255 à chaque fois que l'on fait appel à l'une d'elles. Cela permet d'avoir un ratio compris entre 0 et 1 permettant de caractériser chaque couleur. De plus, l'addition par 128 pour les composantes Cg et Cr permet d'obtenir des octets dont les valeurs varient également entre 0 et 255.

Nous avons vu avec notre encadrant de Valeo que cette méthode de représentation est utilisée dans plusieurs documents de recherche qui ont pour but d'estimer le SpO2. En effet, cette approche s'est démocratisée dans l'étude du problème du SpO2 car l'hémoglobine oxygénée absorbe plutôt la lumière verte tandis que celle désoxygénée absorbe, quant à elle, plutôt la lumière rouge. C'est pour cela que l'utilisation des variables Cg et Cr semble particulièrement adaptée à l'étude de notre problème.

En outre, nous avons pu remarquer dans la littérature scientifique que ce système YCgCr permettait de s'affranchir des conditions de luminosité de la pièce car, comme nous allons le voir par la suite, seul le ratio Cg/Cr est utile dans la prédiction du SpO2 et non pas la luminance. Cela est un atout considérable au vu du jeu de données auquel nous sommes confrontés. En effet, comme nous l'avons expliqué dans la présentation des données, la multiplicité des lieux de collecte des données induit des variations quant aux luminosités de nos images sources.

Voici un schéma représentant l'approche conseillée par Valeo pour traiter ce problème en utilisant le système YCgCr plutôt que le système RGB :

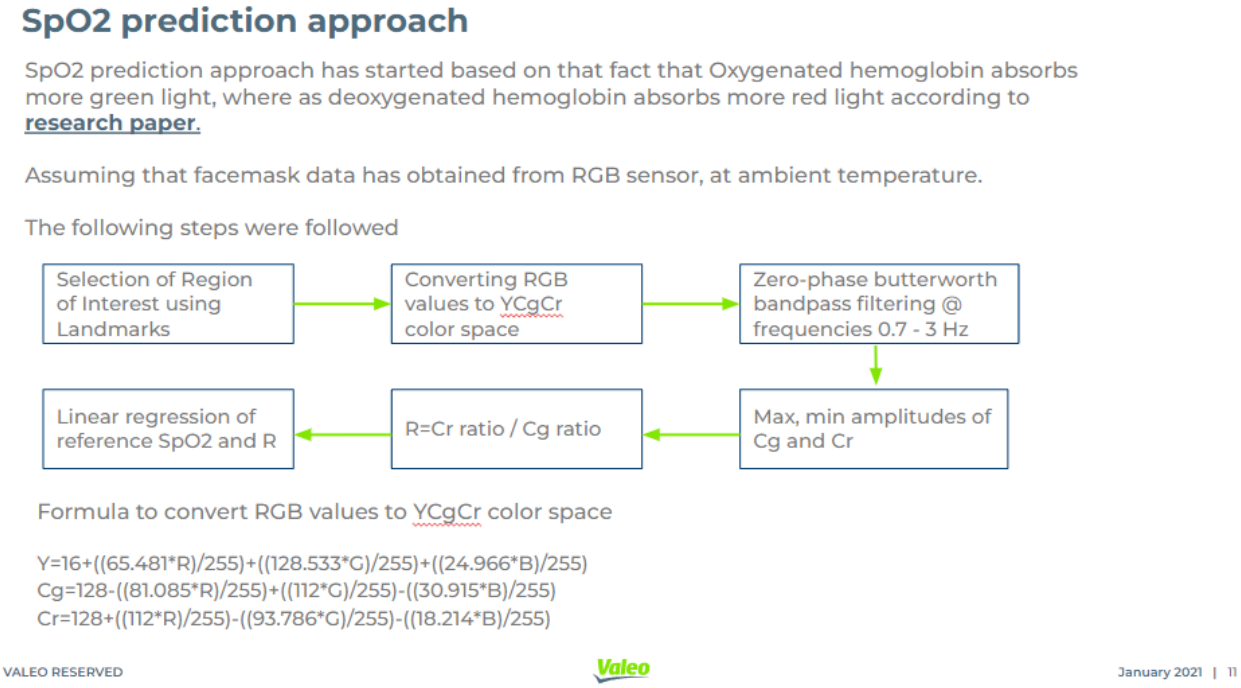


Figure 7 – Pipeline général

Notons toutefois que nous n'avons pas réalisé l'étape de filtrage avec bande passante car cela nécessite de considérer du mouvement (donc une vidéo au lieu d'une image) en entrée. Mais dans ce cas nous ne pourrions plus considérer plusieurs images par personne pour nos algorithmes et nos entraînements mais seulement un signal par personne. Du fait du nombre restreint de données dont nous disposions, nous n'avons donc pas fait cette étape en accord avec notre encadrant chez Valeo.

5.2.1 • Réseaux de neurones ANN (Artificial Neural Networks)

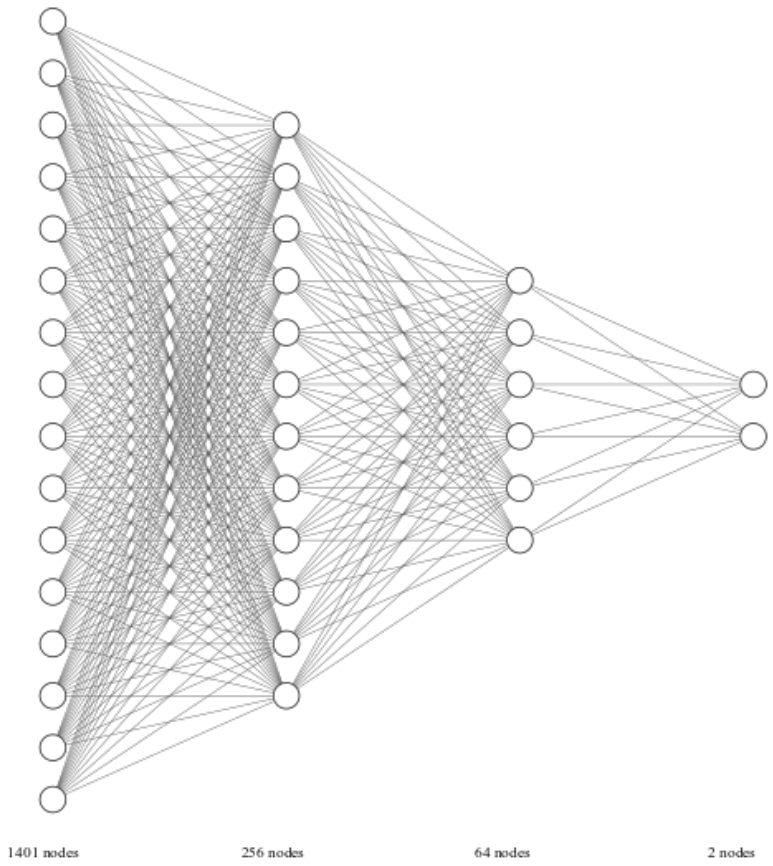


Figure 8 – Architecture de notre réseau de neurones

Le modèle du réseau de neurones artificiel (ANN) est comparable à celui du réseau de neurones du cerveau humain : les neurones ont pour rôle de transmettre l'information et/ou de donner une réponse.

Comme nous pouvons le voir sur l'illustration d'un tel réseau ci-dessus), le réseau est constitué de différents ensembles qui s'emboîtent les uns dans les autres (dans le sens où la sortie/output d'un certain ensemble correspond à l'entrée/input de l'ensemble suivant). Ainsi, un ensemble de neurones reçoit en entrée les sorties de l'ensemble de neurones précédent. Notons que l'ensemble en entrée du réseau de neurones reçoit les informations directement des données du FaceMask et que l'ensemble final renvoie les résultats de l'estimation de la cible.
Nous pouvons remarquer toutefois qu'un réseau de neurones prend en entrée un vecteur, l'image en donnée est donc vue comme un vecteur.

Chaque nœud renvoie une réponse à l'entrée qu'il a reçue et cette réponse est utilisée par le nœud suivant conjointement à toutes les réponses des nœuds de la couche précédente connectés à au nœud suivant en question, pour qu'il puisse se prononcer à son tour. Cela est répété jusqu'à l'ensemble final, assurant une interconnexion de toutes les informations. C'est cette interconnexion qui assure la qualité du modèle et de la réponse finale.

Dans un premier temps se déroule une étape d'apprentissage. L'algorithme reçoit des données d'entraînement. Puis, en fonction des résultats que chaque nœud recevra en entrée de la part des nœuds précédents, il pondérera le poids d'importance qu'il donne à la réponse de chacun de ces nœuds précédents dans la formulation de sa propre réponse.

Ainsi, comme nous l'avons vu, le modèle du réseau de neurones et donc en particulier le modèle d'ANN est particulièrement adapté lorsque les données en entrée sont complexes et peu évidentes à manipuler. Cela est typiquement le cas des FaceMask et beaucoup moins des Raw-Data : ainsi, les modèles que nous avons appliqués à ce second type de données, et en particulier la régression linéaire, sont la preuve de leur facilité d'utilisation dans des algorithmes classiques.

L'ANN possède également d'autres avantages particulièrement utiles : par exemple, la forte interconnexion entre les différents nœuds assure que, quand bien même un nœud est défaillant, l'algorithme peut continuer de fonctionner avec un bon niveau de résultat même si cela affecte certainement la fiabilité du réseau de manière négative. De plus, le réseau complet peut, en repérant des répétitions gagnantes (patterns), développer des connexions extrêmement efficaces. Toutefois, il faut faire attention à ce que de tels patterns soient de réelles combinaisons gagnantes et ne soient pas en réalité le fruit d'une coïncidence chanceuse dans les données d'apprentissage (ou plutôt malchanceuse dans ce cas car cela risquerait de faire s'effondrer la fiabilité des estimations du réseau) qui pourrait ne pas se répéter dans les données que l'on souhaite étudier par la suite. C'est pour cette raison que les biais de classe présents dans nos données sont très regrettables.

5.2.2 • Réseaux CNN (Convolutional Neural Networks)

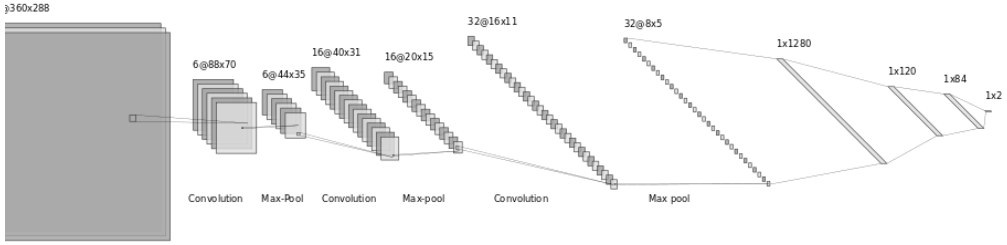


Figure 9 – Architecture de notre CNN

Lorsque l'on manipule des images de données comme pour les FaceMask, nous avons vu précédemment que l'utilisation de réseaux de neurones était particulièrement adaptée.
Étant semblable au réseau de connexion entre les neurones du cortex visuel, le réseau de neurones convolutif semble donc être un bon modèle candidat, d'autant plus que ce modèle de réseau de neurones a été conçu pour traiter des problèmes de machine learning avec des images comme données en entrée.

Notre image du FaceMask reçue en donnée est pavée en différentes zones, chacune étant traitée par un neurone. Puis, de manière similaire à un réseau de neurones artificiel, l'interconnexion entre les neurones assure un bon recoupement des informations et permet à l'algorithme d'effectuer sa prédiction de la valeur cible.

Toutefois, à l'inverse de l'ANN, un noyau de convolution va agir sur l'image d'entrée afin d'en étudier une caractéristique. Un noyau de convolution est une matrice qui agit sur la matrice représentant notre image. De manière métaphorique, nous pouvons donc comparer un noyau de convolution à un filtre qui s'intéresse à une caractéristique de l'image en donnée. Ainsi, plutôt que de voir des nœuds comme pour un réseau ANN, le réseau CNN se modélise plutôt par des "filtres" à appliquer à l'image reçue et qui renvoie donc une nouvelle image. Ces "filtres" forment des couches qui correspondent aux ensembles décrits dans la partie sur l'ANN.

Ainsi, une couche du réseau convolutif reçoit une image en entrée et renvoie une image intermédiaire en sortie. L'image intermédiaire est le résultat de la modification de l'image en entrée par l'application de la couche de "filtres".

À la fin de ce réseau de couches et de la transformation des images données en entrée par les différentes couches de filtres du réseau se déroule une étape d'aplatissement (flattening) qui consiste à transformer les images en sortie de la dernière couche en un (long) vecteur. En effet, il ne faut pas oublier qu'une image est codée par un ensemble de pixels qui correspondent à des matrices de nombres. Construire le vecteur aplati consiste donc simplement à mettre bout à bout les lignes de ces pixels des différentes images en sortie.

À la suite de cette étape de flattening se trouve un réseau de neurones artificiel, dont nous avons déjà décrit le principe dans la section précédente. Comme un réseau de neurones ne prend en compte qu'un vecteur en entrée, l'étape de flattening est indispensable lorsqu'il y a plusieurs images en données d'entrée du réseau de neurones artificiel.

Nous avons déjà détaillé précédemment la phase d'apprentissage pour un ANN ; celui inclus dans le CNN apprend de la même manière. En revanche, pour la partie des couches de filtres, les filtres sont modifiés au fur et à mesure de l'apprentissage afin d'améliorer les résultats des prédictions du réseau.

Ce modèle présente les mêmes avantages et défauts que l'ANN. En revanche, l'utilisation de couches de filtres permet d'obtenir une meilleure précision dans les estimations. De plus,

l'utilisation de ces filtres permet à un réseau CNN de mieux s'adapter à des positionnements différents, différences que nous avons relevées dans notre analyse des données et qui sont principalement dues au fait que Valeo a utilisé plusieurs sites de collectes de données. Cependant, cette adaptation n'est pas possible pour un réseau de neurones artificiel qui, puisqu'il reçoit en entrée un vecteur numérique, a plus de difficultés à se représenter un changement de luminosité ou de point de vue.

5.2.3 • Modèle pour reconnaître les visages

Comme nous l'avons expliqué dans la partie sur la présentation des données FaceMask, l'abondance de données associées à des visages brouillés ou incomplets (voir figure 16 en annexe) nous empêchait d'avoir un set de données suffisant pour travailler correctement avec ces algorithmes. Une étape de tri était donc nécessaire pour la poursuite de notre travail.

Nous avons tout d'abord commencé à trier les masques manuellement avant de nous rendre compte que cette tâche serait non seulement extrêmement fastidieuse mais aussi infiniment longue pour des êtres humains, sachant qu'il y avait 75 645 masques à trier.

Nous nous sommes donc penchés sur l'implémentation d'un algorithme de machine learning capable de trier les données des masques en reconnaissant si elles forment un visage crédible ou non. Pour faire cela, l'algorithme devait non seulement tenir compte des positions des points, mais également des landmarks associés. En effet, nous nous sommes rendu compte qu'un visage peut être complet dans le sens où les points forment bien un visage, mais les landmarks correspondants aux différentes parties du visage (yeux, lèvres par exemples) sont mal placés (voir figure ci-dessous).

Il faut donc trouver un algorithme qui détermine si un FaceMask est non seulement complet au niveau des points, mais aussi juste au niveau des landmarks. Il faut donc se servir des couleurs pour mettre en évidence les landmarks associés aux yeux et aux lèvres.

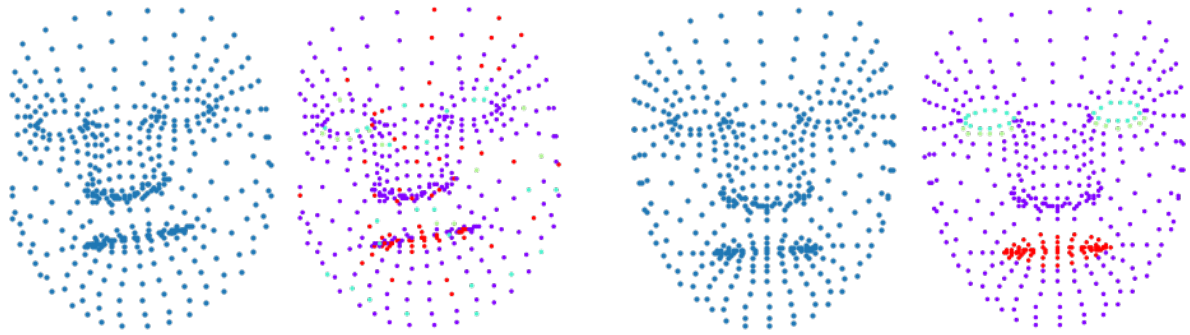


Figure 10 – À gauche, un FaceMask inexploitable - À droite, un FaceMask complet

Nous avons opté pour l'algorithme LeNet, qui est un algorithme de réseau neuronal convolutif introduit en 1998 par Yann Le Cun, l'un des pionniers du Deep Learning. Cet algorithme fut développé dans le but de reconnaître des caractères, il est parfaitement adapté à la reconnaissance d'image, d'où notre choix de l'utiliser.

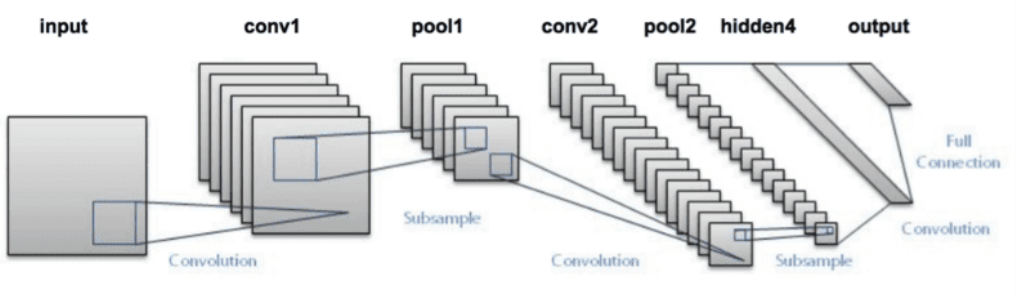


Figure 11 – Architecture de notre CNN

Il s'agit donc d'un algorithme CNN relativement simple, aisé à comprendre et à manipuler. Il est également suffisamment performant pour proposer des résultats très intéressants et satisfaisants.

Cet algorithme comprend deux phases de convolution ainsi que deux phases de pooling. Nous avons vu précédemment que la phase de convolution correspondait à appliquer un "filtre" à l'image reçue en entrée. La phase de pooling consiste, quant à elle, à réduire la taille de l'image tout en généralisant les features extraites de la phase de convolution, en effectuant la moyenne des features dans une certaine zone afin de la réduire à un seul point ensuite par exemple. Cela permet d'avoir une image réduite après le pooling que l'on peut ensuite utiliser pour la seconde phase de convolution et pour la phase de résultats. Le pooling permet donc, par la réduction de la taille de l'image, un traitement plus efficace sans perdre les caractéristiques importantes de l'image.

6
PRÉSENTATION ET COMMENTAIRE DES RÉSULTATS

6.1 Résultats sur les RawData

6.1.1 • La régression linéaire

L'objectif de notre modèle de régression linéaire n'était pas d'obtenir une estimation parfaite de la cible puisque le modèle est bien trop grossier pour cela. En revanche, cela nous a permis de mettre en évidence les relations entre les paramètres des données et la valeur du SpO2, comme nous l'avons vu dans la partie dédiée à cet algorithme.

Nous avons pu mettre en évidence l'importance des données caractérisant le genre, l'âge et la température pour travailler sur ce problème. Cela nous a été très utile lors de notre étape de nettoyage des données. En effet, grâce à cette première estimation, certes grossière, nous savions quels paramètres étaient indispensables dans notre set de données et de quels paramètres nous pouvions nous passer. Cela facilite le tri des données fournies par Valeo et les collectes de données à venir.

De même, comme noté dans la partie sur le nettoyage des données RawData, la visualisation de la matrice de corrélation nous a permis de construire notre propre distance entre données, qui est une adaptation de la distance euclidienne à la lumière de ces corrélations. Nous avons également vu trivialement qu'il n'y avait aucun intérêt à conserver plusieurs colonnes pour les genres et avons donc décidé de les fusionner en une unique colonne "gender".

Ainsi, bien que simpliste, l'algorithme de régression linéaire que nous avons mis en place a répondu à nos attentes, même si ce n'est pas celui que nous utiliserons pour prédire le SpO2. Il nous sert tout de même comme premier point d'ancrage, en offrant une première estimation grossière, et cette modélisation nous a surtout été d'une grande aide dans le traitement des données que nous recevions au fur et à mesure des différents centres de tests.

6.1.2 • Bayes naïf

Comme nous l'avons précisé à la fin de la partie consacrée à ce modèle, en accordant une marge d'erreur de 2% le classificateur bayésien naïf a un taux de précision de 86%. Cependant, du fait des données à notre disposition, ce résultat n'est pas significatif : les personnes étant toutes similaires (jeunes et en bonne santé), le dataset présente un biais considérable. C'est pourquoi l'algorithme détecte très mal les cas à faible SpO2. Pourtant, ce sont des cas

importants car notre système doit servir à aider les services hospitaliers. Il ne peut donc pas se tromper dans ce sens puisqu'il mettrait alors la vie de personnes en danger. Mais, avec les données qui lui sont présentées, puisque l'estimateur bayésien naïf s'appuie sur le maximum de vraisemblance afin de donner l'estimation la plus probable, il est normal qu'il soit lourdement affecté par le grand biais de classe. Les réponses à bas SpO2 relèvent certes de l'exception dans la vie quotidienne, toutefois pour le cadre d'utilisation voulu, la proportion de personnes ayant un faible SpO2 doit être bien supérieure. Avec les données à sa disposition, l'algorithme sait qu'il a peu de chances d'avoir une estimation correcte s'il donne un SpO2 trop faible, et cela quels que soient les paramètres en entrée. C'est pourquoi il répond presque systématiquement en donnant les valeurs de SpO2 principales contenues dans le set de données, c'est-à-dire 97 et 98.

Notons que tous nos algorithmes de prédiction du SpO2 sont affectés par cet important biais de classe et qu'il serait intéressant de les tester avec un dataset plus équilibré. Nos accompagnateurs chez Valeo nous ont répondu qu'il était possible que l'on reçoive ce set bientôt (à l'horizon de quelques semaines ou quelques mois), à savoir des données plus diversifiées issues de patients en moins bonne santé, notamment grâce au partenariat avec le CHU Henri Mondor. Nous nous tenons évidemment à leur disposition pour cette occasion.

6.1.3 • Random Forests

Par rapport au modèle de classification bayésien naïf, le modèle de forêts aléatoires présente une amélioration majeure : il peut détecter des SpO2 faibles même si, là encore, une bonne prédiction d'un SpO2 faible relève plus de l'exception que de la norme.

Nous pouvons observer ci-dessous un graphique de points avec la valeur réelle du SpO2 en abscisse et celle estimée par l'algorithme en ordonnée.

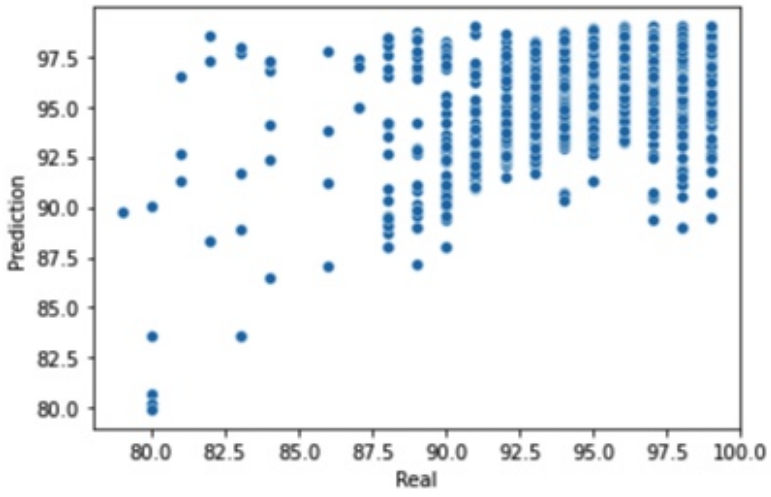


Figure 12 – Prédiction de l'algorithme random forest sur notre dataset

Comme nous pouvons le constater, les résultats de l'algorithme sur notre jeu de données sont loin d'être satisfaisants. Par rapport à l'algorithme précédent, il se trompe davantage concernant les SpO2 élevés.

Cependant, comme expliqué au paragraphe précédent, ces résultats sont à mettre en relief avec le jeu de données à notre disposition : ces données sont peu nombreuses et biaisées, ce qui complique la tâche de l'algorithme notamment pour sa phase d'apprentissage.

6.2 Résultats sur les données FaceMask

6.2.1 • Résultats des modèles ANN et CNN

Nous avons appliqué nos deux algorithmes de réseau de neurones à notre problème avec les données conservées grâce à notre modèle de reconnaissance des visages. Toutefois, avec ces données, les résultats obtenus sont loin d'être probants. En effet, nos modèles semblent répondre de manière aléatoire à ce problème de classification.

Comme nous pouvons le voir sur le graphique ci-dessous, l'algorithme converge vers un taux de réponses correctes de 60% alors que la classe correspondant à un SpO2 strictement supérieur à 97 représente également 60% des données.

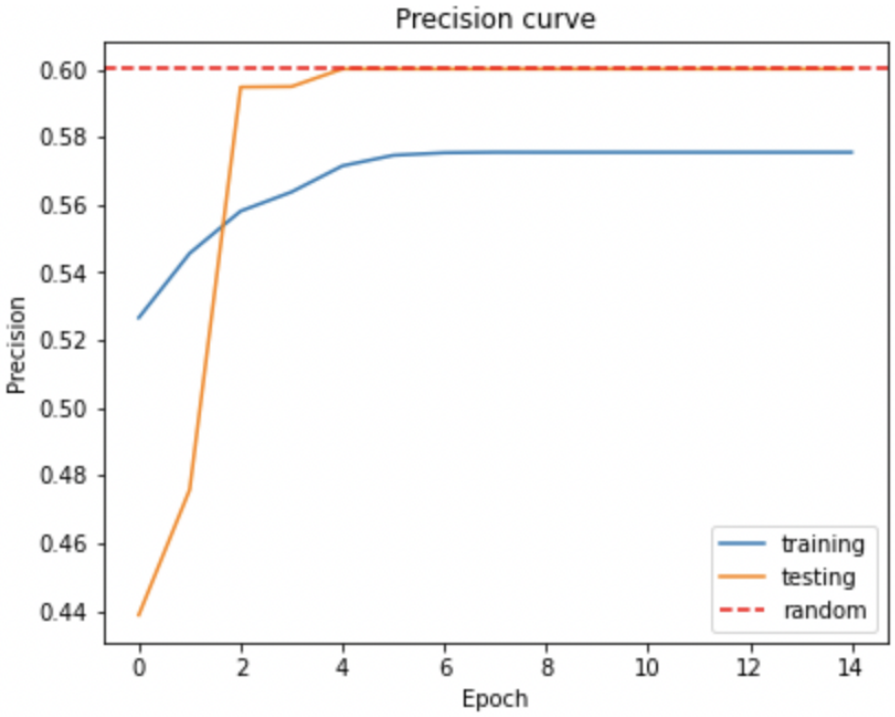


Figure 13 – Prédiction de l'algorithme CNN sur notre dataset

Notons néanmoins que, comme pour les autres algorithmes utilisant notre base de données, ce résultat n'est pas significatif à cause du manque de données exploitables et de son important biais de classe. Avec les futures données collectées par Valeo que nous devrions recevoir bientôt, nous pourrions avoir plus de données et ainsi entraîner nos modèles pour obtenir des résultats plus précis.

6.2.2 • Résultats du modèle pour reconnaître les visages

Puisque les acquisitions prenaient plusieurs images de FaceMask par volontaire, nous disposions de 75 645 masques à trier selon s'ils étaient utilisables ou non. Comme nous l'avons mentionné dans la présentation du modèle utilisé pour traiter ce problème, il était illusoire de songer à tous les trier à la main. En revanche, pour que notre algorithme puisse passer la phase d'entraînement, il lui fallait donner un premier jeu de données que nous aurions préalablement triées. Nous avons donc travaillé manuellement afin d'avoir un jeu d'environ 5000 données. Puis, nous avons testé notre algorithme sur ce jeu de données et avons obtenu un taux de réussite de 97.9%.

Cet excellent résultat nous a permis d'utiliser l'algorithme pour effectuer le tri des données FaceMask utilisables, ce qui a représenté un avancement majeur dans notre projet ainsi qu'un modèle qui sera utile lors de la réception de nouvelles données, particulièrement si nous avons accès à des données du CHU Henri Mondor qui nous permettraient d'avoir un dataset plus diversifié, et donc de travailler à détecter des cas de SpO2 plus faibles, ce qui est la réelle motivation de notre projet.

De plus, nous avons continué de trier manuellement les données FaceMask afin d'avoir un jeu d'entraînement plus complet et d'améliorer la performance de notre algorithme. Avec ces nouvelles données d'entraînement, qui s'élèvent au nombre de 6000 environ, nous avons amélioré l'exactitude du modèle à 99.3%. Le modèle était d'ailleurs devenu suffisamment performant pour détecter des erreurs d'inattention que nous avions commises dans notre tri manuel.

7 CONCLUSION

L'estimation du SpO2 à l'aide d'un dispositif sans contact est une problématique riche en opportunités. Nous avons eu la chance de l'étudier avec des encadrants qui ont su nous aiguiller tout au long de ce projet.

L'entreprise Valeo, qui a proposé ce sujet, nous a permis de travailler sur des données provenant d'acquisitions réalisées dans différents sites à travers le monde.

Une partie importante de notre travail s'est résumée à l'analyse et le tri de ces données. Au cours de cette étape, nous avons pu constater que ces données présentaient un biais important. Ce biais constitue un réel obstacle pour nos algorithmes qui ont des difficultés à prédire correctement le SpO2 (surtout pour des cas qui diffèrent de la majorité des données biaisées). De plus, le faible nombre de données à notre disposition s'avère également être un frein au bon fonctionnement des algorithmes que nous avons mis en place. L'entreprise nous a prévenu qu'elle pourrait bientôt nous donner accès à de nouvelles données qui ne souffriraient pas de la contrainte de ces biais de classe. Nous nous tenons à leur disposition pour travailler avec ces données supplémentaires.

Néanmoins, il convient de noter que nous avons tout de même réussi à réaliser un algorithme de reconnaissance des données FaceMask utilisables qui présente de très bons résultats (plus de 99% de succès) et donne satisfaction.

Ce projet a été pour nous l'occasion de nous plonger dans un véritable projet de data science. Même si nous n'avons finalement pas pu créer un modèle d'estimation du SpO2 satisfaisant du point de vue industriel, les analyses que nous avons pu mener ont été riches en enseignements divers et variés, et les équipes de Valeo en charge du portique santé pourront désormais s'appuyer sur nos résultats pour continuer le développement de leur produit.

8 ANNEXE

	/'FaceMaskLog'/'Count'	/'Landmarks'/'Landmarks'	...	/'T°max(max)'/'T°max(max)'	/'T° brutes'/'T° brutes'
0	468.0	0	...	0.000000	26.279999
1	468.0	1	...	35.181362	20.850006
2	468.0	2	...	35.181362	21.75
3	360.0	3	...	35.181362	23.99002
...	...	...	...	...	...
38373	NaN	464	...	NaN	0.00
38374	NaN	465	...	NaN	0.00
38375	NaN	466	...	NaN	0.00
38376	NaN	999	...	NaN	999

	/'Measured Data'/'HR'	/'Measured Data'/'RR'	...	/'Diag Results'/'RR'	/'Diag Results'/'T'
0	0.0	0.0	...	21.49103	35.181362
1	0.0	0.0	...	NaN	NaN
2	0.0	0.0	...	NaN	NaN
3	0.0	0.0	...	NaN	NaN
...	...	...	...	...	...
75	68.0	22.0	...	NaN	NaN
76	68.0	22.0	...	NaN	NaN
77	67.0	22.0	...	NaN	NaN
78	67.0	22.0	...	NaN	NaN

Figure 14 – Exemple tronqué d'un tableau des données

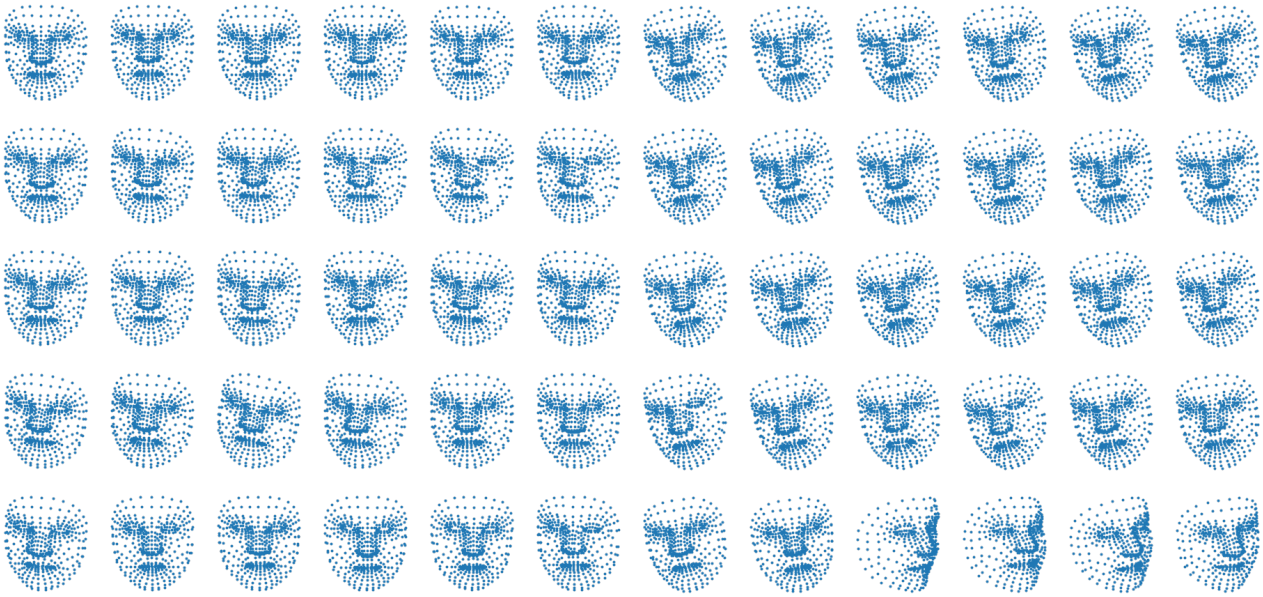


Figure 15 – Représentation des données FaceMask

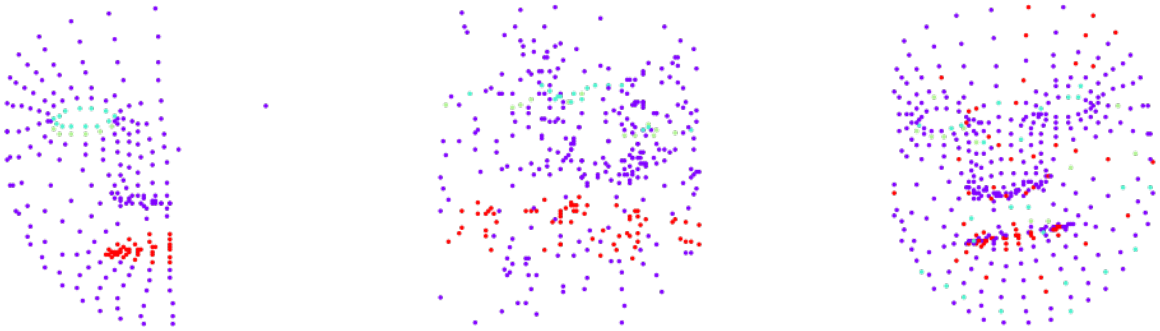


Figure 16 – Représentation de données FaceMask brouillés ou incomplets

Septembre - Octobre	Recherches bibliographiques	Tout le groupe
Octobre - Février	Analyse des données	Tout le groupe
Février - Avril	Modèles RawData	Idriss Ben Abdallah, Marc-Olivier Quéran et Zoulu Ding
Février - Avril	Modèles FaceMask	Ivan Bettannier, Adrien Schneider et Long Vân Tran Ha
Mars - Avril	Reconnaissance des visages	Long Vân Tran Ha

Table 1 – Répartition du travail

RÉFÉRENCES

[1] BARFORD Charlotte [et al.]. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department-a prospective cohort study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 10 avril 2012, Vol.20,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://sjtrem.biomedcentral.com/articles/10.1186/1757-7241-20-28.

[2] BEIDERMAN Yevgeny [et al.]. Demonstration of remote optical measurement configuration that correlates to glucose concentration in blood. *Biomedical Optics Express*, 14 mars 2011, Vol.2, pp. 858–870, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.researchgate.net/publication/51042374_Demonstration_of_remote_optical_measurement_configuration_that_correlates_to_glucose_concentration_in_blood.

[3] GU W [et al.]. *2009. A h-Shirt-Based Body Sensor Network for Cuffless Calibration and Estimation of Arterial Blood Pressure In IEEE Computer Society (éd.), Sixth International Workshop on Wearable and Implantable Body Sensor Networks. Berkeley p 151-155.*, ISBN 978-0-7695-3644-6.

[4] GUAZZI Alessandro [et al.]. Non-contact measurement of oxygen saturation with an rgb camera. *Biomedical Optics Express*, 11 août 2015, Vol.6,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574660/.

[5] JEONG In Cheol and FINKELSTEIN Joseph. Introducing contactless blood pressure assessment using a high-speed video camera. *Journal of Medical Systems*, avril 2016, Vol.40, [en ligne].[consulté le 22 septembre 2021]. Disponible à l'adresse https://pubmed.ncbi.nlm.nih.gov/26791993/.

[6] KAMAL A [et al.]. Skin photoplethysmography-a review. *Computer Methods and Programs in Biomedicine*, Avril 1989, Vol.28, pp. 257-269 [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.sciencedirect.com/science/article/pii/0169260789901594.

[7] KEBE Mamady [et al.]. Human vital signs detection methods and potential using radars : A review. *Sensors*, 6 mars 2020, Vol.20,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7085680/.

[8] KUZMINA I [et al.]. Towards noncontact skin melanoma selection by multispectral imaging analysis. *Journal of Biomedical Optics*, 1er juin 2011, Vol.16, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-16/issue-06/060502/Towards-noncontact-skin-melanoma-selection-by-multispectral-imaging-analysis/10.1117/1.3584846.full.

[9] LEATHAM Aubrey. Phonocardiography. *British Medical Bulletin*, 1952, Vol.8, Issue 4, pp. 333–342,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847938/.

[10] MKENGNE A [et al.]. Spectracam : A new polarized hyperspectral imaging system for repeatable and reproducible in vivo skin quantification of melanin, total hemoglobin, and oxygen saturation. *Skin Research and Technology*, août 2017, Vol.24, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://www.researchgate.net/publication/318889488_SpectraCamR_A_new_polarized_hyperspectral_imaging_system_for_repeatable_and_reproducible_in_vivo_skin_quantification_of_melanin_total_hemoglobin_and_oxygen_saturation`.

[11] MOSTOV K, LIPSTEN E, and BOUTCHKO R. Medical applications of shortwave fm radar : Remote monitoring of cardiac and respiratory motion. *Medical Physics*, Mars 2010, Vol.8,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847938/`.

[12] NA HYE Kim [et al.]. Non-contact oxygen saturation measurement using ycgcr color space with an rgb camera. *Sensors*, 12 septembre 2021, Vol.21,[en ligne].[consulté le 9 mars 2022]. Disponible à l'adresse `https://www.mdpi.com/1424-8220/21/18/6120`.

[13] NEWTON D [et al.]. Comparison of macro and micro-lightguide spectrophotometric measurements of microvascular hemoglobin oxygenation in the tuberculin reaction in normal human skin. *Physiological Measurement*, Mai 1994, Vol.15, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://pubmed.ncbi.nlm.nih.gov/8081190/`.

[14] SHENHAV Asaf [et al.]. Optical sensor for remote estimation of alcohol concentration in blood stream. *Optics Communications*, 15 février 2013, Vol.289, pp. 149-157, [en ligne].[consulté le 22 septembre 2021]. Disponible à l'adresse `https://www.sciencedirect.com/science/article/pii/S0030401812011248`.

[15] STRATONNIKOV A and LOSCHENOV V. Evaluation of blood oxygen saturation in vivo from diffuse reflectance spectra. *Journal of Biomedical Optics*, Octobre 2001, Vol.6, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://pubmed.ncbi.nlm.nih.gov/11728206/`.

[16] SUN Zhiyuan [et al.]. Robust non-contact peripheral oxygenation saturation measurement using smartphone-enabled imaging photoplethysmography. *Biomedical Optics Express*, 1^{er} mars 2021, Vol.12, pp. 1746–1760 [en ligne].[consulté le 22 septembre 2021]. Disponible à l'adresse `https://www.osapublishing.org/boe/fulltext.cfm?uri=boe-12-3-1746&id=448656`.

[17] TSUMURA N, HANEISHI H, and MIYAKE Y. Independent-component analysis of skin color image. *Journal of the Optical Society of America*, 1^{er} septembre 1999, Vol.16, [en ligne].[consulté le 22 septembre 2021]. Disponible à l'adresse `https://www.osapublishing.org/josaa/fulltext.cfm?uri=josaa-16-9-2169&id=1267`.

[18] VAN GASTEL Mark, STUIJK Sander, and DE HAAN Gerard. New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring. *Scientific Reports*, 7 décemblre 2016, Vol.6, [en ligne].[consulté le 22 septembre 2021]. Disponible à l'adresse `https://www.nature.com/articles/srep38609`.

[19] WANG X and SHAO D. 2022. human physiology and contactless vital signs monitoring using camera and wireless signals in academic press (éd.), contactless vital signs monitoring. p 1-24. ISBN 978-0-12-822281-2.

[20] WILL Christoph [et al.]. Radar-based heart sound detection. *Scientific Reports*, 26 juillet 2018, Vol.8,[en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6070547/`.

[21] YAMAKOSHI Yasuhiro, OGAWA Mitsuhiro, and TAMURA Toshiyo. Multivariate regression and classification models for estimation of blood glucose levels using a new non-invasive optical measurement technique named "pulse-glucometry. *The Open Optics Journal*, Septembre 2009, Vol.3, pp. 63-69, [en ligne].[consulté le 15 septembre 2021]. Disponible à l'adresse `https://www.researchgate.net/publication/228652721_Multivariate_Regression_and_Classification_Models_for_Estimation_of_Blood_Glucose_Levels_Using_a_New_Non-invasive_Optical_Measurement_Technique_Named_Pulse-Glucometry`.