

Rapport de projet AutoML

Axel Roy

juil. 22, 2017

Table des matières

| | | |
|----------|--|-----------|
| 1 | Introduction | 3 |
| 1.1 | Contexte du projet | 3 |
| 1.1.1 | Human Brain projet | 3 |
| 1.1.2 | Présentation de la plateforme MIP | 3 |
| 1.1.3 | But du projet | 7 |
| 2 | Cahier des charges (lien vers les annexes je suppose, en sachant qu'il est expliqué en détail dans le document) | 8 |
| 3 | Etat de l'Art | 9 |
| 3.1 | Théorie Machine Learning | 9 |
| 3.1.1 | Apprentissage supervisé | 10 |
| 3.1.2 | Apprentissage non supervisé | 10 |
| 3.1.3 | Apprentissage semi-supervisé | 11 |
| 3.2 | Optimisation automatique du pipeline d'apprentissage | 11 |
| 3.3 | Technologies | 13 |
| 3.3.1 | TPOT | 13 |
| 3.3.2 | Systèmes distribués | 13 |
| 3.3.3 | Mesos | 15 |
| 3.3.4 | Marathon | 16 |
| 3.3.5 | Chronos | 16 |
| 3.3.6 | Docker | 17 |
| 3.3.7 | Scala | 19 |
| 3.3.8 | AKKA | 19 |
| 4 | Analyse | 20 |
| 4.1 | Woken | 20 |
| 4.1.1 | Place de Woken dans l'architecture | 20 |
| 4.1.2 | Fonctionnement interne de Woken | 22 |
| 4.1.3 | Fonctionnement actuel des containers Docker | 25 |
| 4.1.4 | Tests préliminaires avec TPOT | 26 |
| 4.1.5 | Le cas de Marathon | 26 |

| | | |
|-----------|---|-----------|
| 5 | Conception | 27 |
| 5.1 | Système de containers interactifs | 27 |
| 5.2 | Modification globale du workflow Woken | 27 |
| 5.3 | Nouveau diagramme d'acteurs imaginé, et comment on coupe le workflow actuel | 29 |
| 6 | Implémentation réalisée | 30 |
| 6.1 | Création d'un container interactif | 30 |
| 6.1.1 | Problème initial | 30 |
| 6.1.2 | Présentation des solutions au problème | 30 |
| 6.1.3 | Choix effectué | 30 |
| 6.2 | Modification du workflow Woken | 30 |
| 6.2.1 | Ajout du nouveau container dans la configuration | 30 |
| 6.3 | Intégration de TPOT | 30 |
| 6.3.1 | A déterminer, mais je suppose : Les contraintes posées par la biblio- thèque, les choix qui ont du être effectués. | 30 |
| 6.4 | Eventuellement, si plus de travail a été effectué, présentation de celui-ci. . . . | 30 |
| 7 | Validation (Expérience) | 31 |
| 8 | Conclusion | 32 |
| 8.1 | Etat des lieux au moment du rendu | 32 |
| 8.2 | Perspectives et améliorations | 32 |
| 8.3 | Bilan personnel (Présenter ce qui apporte quelque chose) | 32 |
| 9 | Remerciements | 33 |
| 10 | Annexes, références et Table des illustrations. | 34 |

Le présent document fait office de rapport de projet. Il permet de comprendre le contexte de celui-ci, de reconstituer son cheminement du projet, de comprendre les choix et les déductions effectuées, ainsi que de connaître l'état final du travail et les perspectives d'amélioration.

1.1 Contexte du projet

Le présent projet s'inscrit dans le cadre du travail de Bachelor en Informatique option « Développement logiciel et multimédia », réalisé à la HE-ARC de Neuchâtel.

Le projet est effectué pour le CHUV-LREN dans le cadre du projet Human Brain Project.

1.1.1 Human Brain projet

Ce projet s'inscrit dans le cadre du projet Européen « Human Brain Project ». Ce chapitre vise à expliquer le contexte de la partie du projet qui nous intéresse.

1.1.2 Présentation de la plateforme MIP

Le but du sous-projet 8 du HBP est de fournir une plateforme pour effectuer des expériences neuroscientifiques sur des données de patients recueillies à travers les cliniques et hôpitaux partenaires. Etant donné la nature médicale de ces données, elles sont bien évidemment anonymisées, et il n'est pas possible de retrouver les données d'un patient, car les données sont présentées sous la forme d'agrégation par caractéristique, comme le présente la [figure 1.1](#)

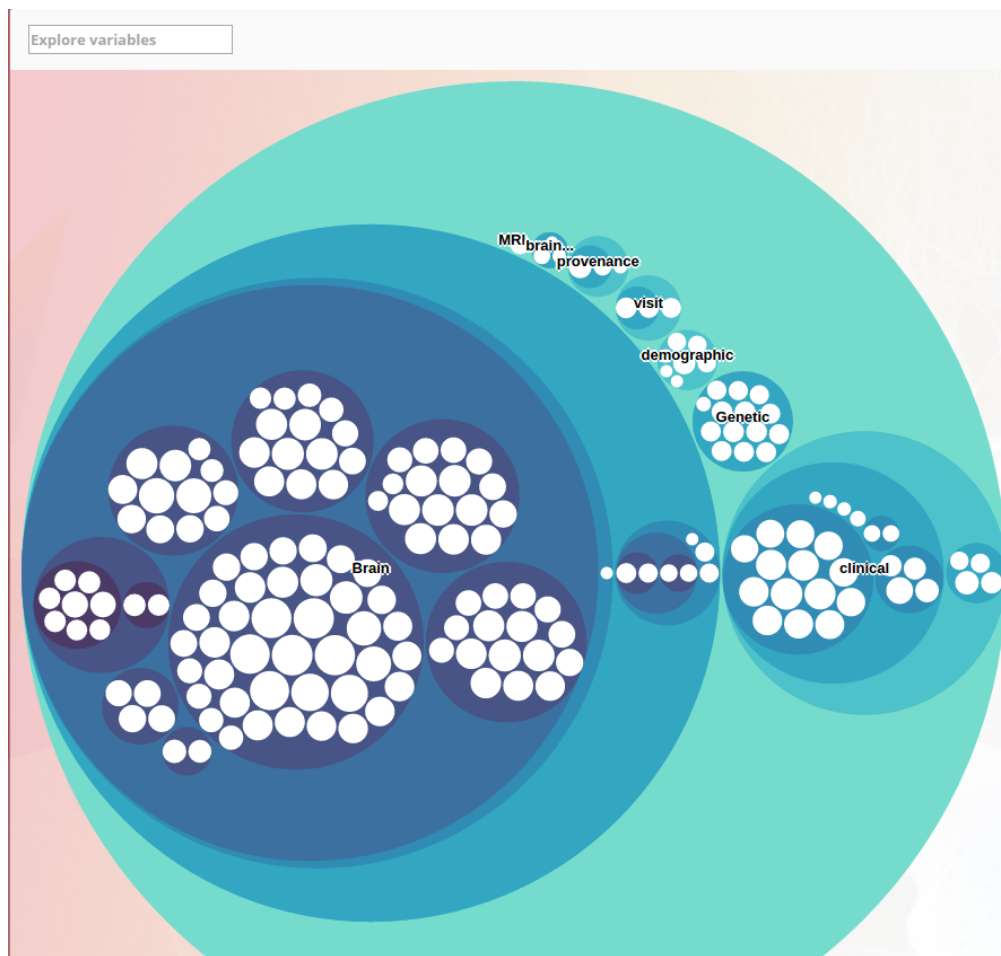


Fig. 1.1 : *Représentation des caractéristiques d'intérêts.*

En sélectionnant un des ronds blancs, on accède à la variable en question, et on peut observer différentes statistiques, comme par exemple des vues sous forme d'histogrammes [cf [figure 1.2](#)]

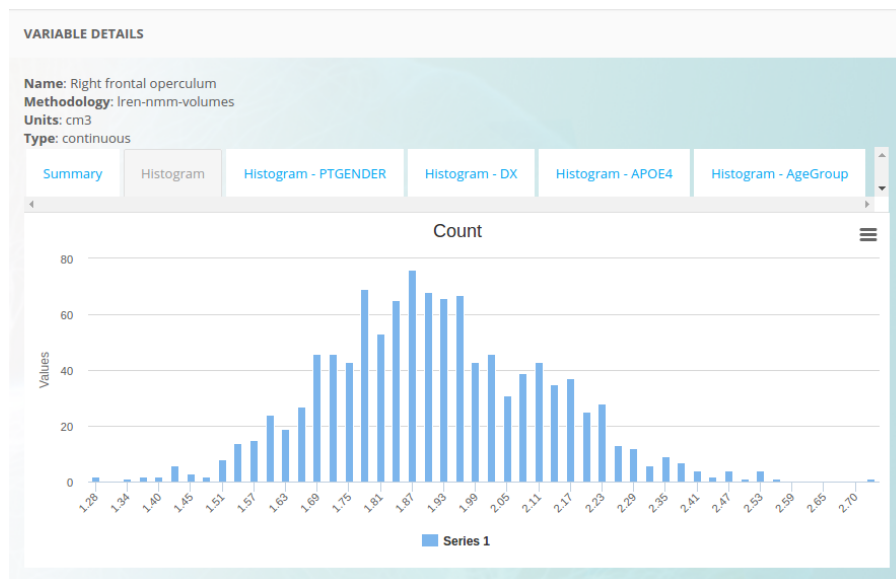


Fig. 1.2 : Exemple d'histogramme d'une variable.

Il est ainsi possible d'accéder à toutes les caractéristiques médicales et ainsi de les analyser manuellement. La plateforme permet aussi de formuler des expériences basées sur les données, afin de proposer un modèle personnalisé qui permet d'essayer de trouver des liens entre les variables des patients et leur diagnostics médicaux. La plateforme permet aussi de formuler des expériences liées à Alzheimer, mais d'autres maladies neurologiques pourraient être visées. À partir d'une caractéristique, l'utilisateur peut décider de formuler une expérience en choisissant dans laquelle des catégories suivantes il compte l'impliquer :

- Variable
- Co-variable
- Filtre

Via l'interface suivante présentée en figure 1.3.

| ADD ALL AS VARIABLE | ADD ALL AS COVARIABLE | | ADD ALL AS FILTER |
|--|---|--|-------------------|
| VARIABLE <input checked="" type="checkbox"/> Left cuneus | NOMINAL <input checked="" type="checkbox"/> Sex | CONTINUOUS <input checked="" type="checkbox"/> Age | FILTERS |

Fig. 1.3 : Exemple de formulation d'expérience, étape sélection des variables. Cet exemple vise à trouver un lien entre la quantité de matière grise dans le Cuneus en fonction de l'âge et du sexe.

Ce qui nous amène vers la possibilité d'analyser des graphes mêlant les différentes variables. Il est encore possible de paramétrer la représentation sur l'axe via une boîte à outils, afin de faire ressortir les informations intéressantes, comme présenté à la figure 1.4

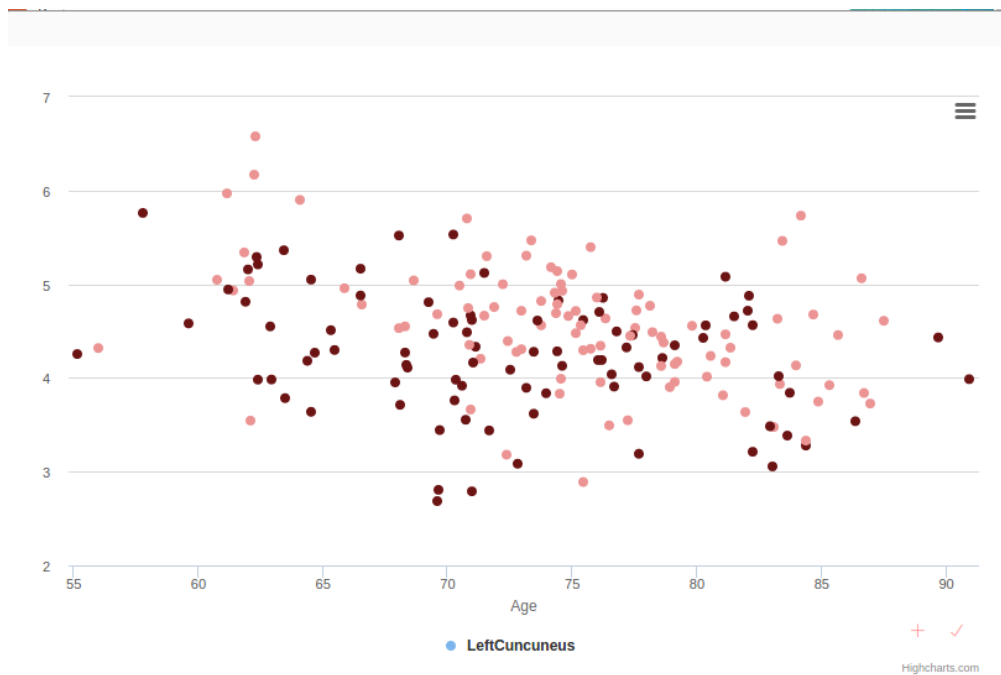


Fig. 1.4 : Résultat de l'expérience formulée à la :num: 'figure #variables'. Représentation de la quantité de matière grise en cm3 en fonction de l'âge et du sexe (bordeau = femme, rose = homme).

La partie intéressante dans le cadre de ce projet est la possibilité, à partir des variables sélectionnées, de lancer une expérience d'apprentissage automatique (Machine Learning) afin de trouver le modèle qui permet de représenter au mieux le lien entre les caractéristiques et le diagnostique.

L'aide pour la configuration de l'expérience est présentée comme en [figure 1.5](#)

About Running Experiments

An experiment is a set of algorithm(s) and their configuration applied to the variables already selected. You may choose same algorithms more than once providing that you change the configuration parameters.

You can design your own MIP Experiment by doing the following:

1. Select a Machine Learning method on the left
2. If required, configure parameters (e.g. "k")
3. Add to Experiment
4. Optionally operate the same steps with other algorithms and/or parameters
5. Give a name to the Experiment
6. Select your k-fold Validation (to be applied to predictive model algorithms)
7. Run Experiment

Wait for results.

Fig. 1.5 : Aide pour la formulation d'une expérience de Machine Learning.

Les étapes 1 et 2 sont celles qui nous intéressent :

L'étape 1 correspond à la sélection d'un algorithme de *Machine Learning* dans la liste fournie (catégories : analyse statistique, extraction de caractéristiques et modèle prédictif). Le modèle choisi influence fortement les résultats de l'expérience.

Lorsque le modèle est sélectionné, il est possible, suivant le modèle, de devoir renseigner des

« **paramètres** » pour celui-ci. Nous appellerons ces paramètres des « **hyper-paramètres** », afin d'éviter la confusion avec les paramètres qui sont les coefficients internes qui ont été déterminés après l'entraînement. Les hyper-paramètres définissent un fonctionnement interne (par exemple, pour le modèle KNN, l'hyper-paramètre k désigne le nombre des voisins les plus proches sur lesquels on veut travailler). Le choix de ces hyper-paramètres est donné au points deux de cette marche à suivre. Pour un même modèle, le choix d'un hyper-paramètres plutôt qu'un autre change à nouveau drastiquement les résultats.

Il peut définir plusieurs configurations « modèle-paramètres » pour une expérience. Une expérience ne donne pas instantanément ses résultats. L'utilisateur est notifié lorsque les résultats sont consultables.

C'est ici que s'inscrit le projet. L'utilisateur, qui est probablement plus un spécialiste en neurosciences qu'en informatique, se trouve obligé de paramétrer et choisir des données qui sont liées uniquement à l'informatique.

1.1.3 But du projet

Ce projet a pour but de mettre en place un moyen pour que l'utilisateur n'ait plus à s'occuper du choix du modèle et du paramétrage pour son expérience, et que la plateforme s'occupe de trouver automatiquement la meilleure configuration possible. Dans l'idéal, l'utilisateur n'a qu'un bouton à presser pour cette étape.

Cahier des charges (lien vers les annexes je suppose, en sachant qu'il est expliqué en détail dans le document)

Se référer au cahier des charges fourni en annexes.

Avant de se lancer dans la partie plus en détail dans la description de la plateforme, il est intéressant d'effectuer un état de l'art des technologies qui pourraient nous intéresser. Etant donné que le projet consiste à ajouter des fonctionnalités à un projet existant, cette section décrira les technologies actuellement existantes, ainsi que les technologies ajoutées, ou tout du moins leur champ d'application.

Cette section est rédigée en listant les différentes technologies, de la plus globale à la plus précise en terme d'utilisation dans le projet.

3.1 Théorie Machine Learning

Le *Machine Learning* (apprentissage automatique en français), est un champ d'activité de l'intelligence artificielle qui vise à permettre à une machine d'apprendre par elle-même plutôt que d'en fixer tous les comportements de manière programmatique. Elle est particulièrement utilisée dans les problématiques où le nombre de cas est trop important pour être codés à la mano. Le panel d'utilisation est large, il peut par exemple concerner :

- L'analyse de graphes ou de données
- La classification d'individus
- La résolution de problèmes de régression
- La reconnaissance d'objets
- L'analyse de documents (notamment pour les moteurs de recherche)
- La reconnaissance de caractères manuscrits
- L'aide au diagnostics médicaux

Dans notre cas, l'apprentissage automatique est implémenté dans la plateforme via les méthodes suivantes :

- Résumé statistique ;
- Analyse de la variance (anova) ;

- Régression linéaire ;
- KNN
- Classification naïve bayésienne

Mais on peut aussi ajouter à la plateforme d'autres méthodes d'apprentissage automatique via des containers Docker vierges qui sont fournis par le projet.

3.1.1 Apprentissage supervisé

Dans cette méthodologie, on connaît déjà les classes que l'on souhaite pouvoir déterminer automatiquement via l'algorithme. Ces classes sont tirées des données par un expert. Dans certains cas, il est aussi possible d'attribuer une probabilité d'appartenance à une classe. L'apprentissage se déroule généralement en deux phases. La première phase est dite d'entraînement. Elle consiste à déterminer un modèle qui permet de reproduire pour de nouvelles données la même classification/régression que celle donnée via les labels. La seconde phase est dite de validation. Elle consiste à déterminer si le modèle entraîné est pertinent, via des méthodes métriques. Ces deux phases ne s'effectuent pas sur les mêmes données. La phase d'entraînement nécessite une quantité d'informations suffisantes afin d'avoir un modèle représentatif.

3.1.2 Apprentissage non supervisé

Cet apprentissage s'applique à des données qui ne sont pas labellées par des classes. C'est ici à la machine de déterminer les différentes classes qui représentent le problème. A partir d'un ensemble de données en entrées, il va chercher à créer des classes représentatives pour celles-ci, en maximisant la distance inter-classe, et en minimisant la distance des éléments intra-classe comme représenté sur la [figure 3.1](#).

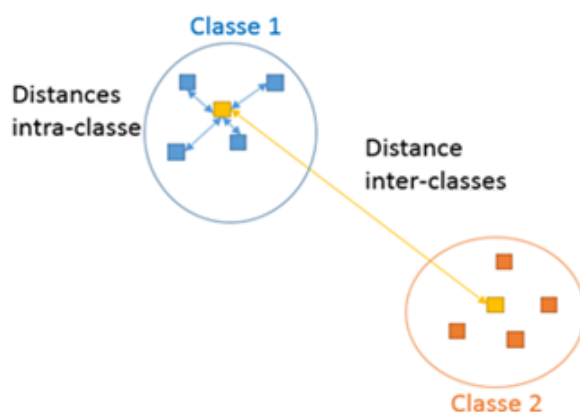


Fig. 3.1 : Représentation des distances inter-classe et intra-classe. Illustration issue du site MSDN [1]

Cette méthodologie peut aussi permettre d'analyser la relation entre les variables, par exemple pour réduire la dimension des vecteurs d'entrées.

3.1.3 Apprentissage semi-supervisé

Etant donné que l'apprentissage supervisé nécessite une labélisation des données par expert, il devient très coûteux de réaliser ce travail au fur et à mesure que les données augmentent. L'utilisation de données labellées, liées à des données non labellées, peut permettre d'améliorer la qualité de l'apprentissage. Par exemple, il est ainsi possible d'utiliser un classificateur créé par l'apprentissage supervisé, et un autre créé par l'apprentissage non-supervisé.

Idéalement, les deux classificateurs ne se basent pas sur les mêmes caractéristiques, ce qui permet de recouper les deux classificateurs afin d'affiner la classification finale.

3.2 Optimisation automatique du pipeline d'apprentissage

De manière générale, le *Machine Learning* est décrit comme une suite d'opérations à effectuer de manière séquentielle pour permettre de résoudre une problématique. On parle dès lors de pipeline, étant donné que chaque étape est effectuée, à la manière d'un flux d'opérations, de la première à la dernière.

Ce pipeline est généralement découpé en deux phases distinctes :

- Extraction, normalisation et éventuellement construction des caractéristiques à partir des données brutes.
- Application d'un modèle statistique ou linéaire pour effectuer, selon la problématique, une classification ou une régression.

On peut représenter ce flux via la [figure 3.2](#) : On peut en décrire les phases ainsi :

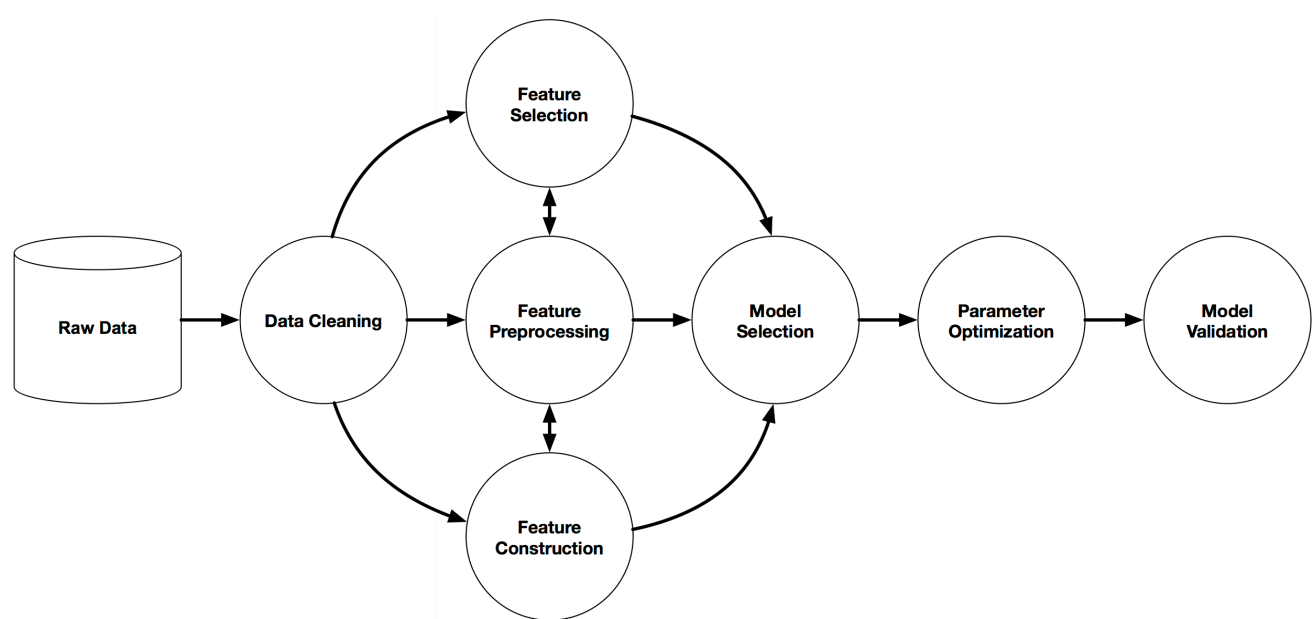


Fig. 3.2 : Exemple d'un pipeline de Machine Learning, tiré de la documentation TPOT [2] et adapté pour supprimer les parties liées à TPOT.

- **Data Cleaning** : Mise en forme des données et nettoyage. Ceci peut consister à renseigner les données manquantes.
- **Features Preprocessing** : Transformation des caractéristiques pour les rendre plus utilisables dans le contexte, par exemple en les normalisant.
- **Features Selection** : Sélection des caractéristiques les plus pertinentes pour le modèle.
- **Feature Construction** : Création de nouvelles caractéristiques à partir des données.
- **Model Selection** : Sélection du type de modèle ainsi que les hyper-paramètres liés à celui-ci (p.e. pour un réseau de neurones, le nombre de couches de neurones). Actuellement, l'utilisateur doit les configurer lui-même, et même un utilisateur expert ne peut pas garantir que ce sont les meilleurs hyper-paramètres possibles.
- **Parameter Optimization** : le choix d'un modèle détermine les paramètres qui lui sont liés (p.e. pour un réseau de neurones, le poids de chaque neurone). Ces paramètres influencent énormément la performance du modèle. Ils sont optimisés lors de cette phase.
- **Model Validation** : En sortie, nous avons, pour un ensemble de caractéristiques donnée, un modèle et le hyper-paramètres de ce modèle. Il faut ensuite valider ce modèle sur un ensemble de sujets différents afin de déterminer sa pertinence.

Dans une approche traditionnelle d'optimisation d'une expérience de *Machine Learning*, on essaie de faire varier les hyper-paramètres du modèle (p.e via les grid-search [3] de Scikit-Learn [4]).

Cette méthode permet d'optimiser les hyper-paramètres du modèle. Ce dernier doit avoir été sélectionné manuellement auparavant par l'utilisateur. De plus, l'étendue et le pas des hyper-paramètres sont eux-aussi déterminés manuellement. Cela réduit le domaine d'exploration.

Une tendance émergente de ces dernières années est d'utiliser des méthodes d'intelligence artificielle pour explorer l'espace des solutions de manière automatique et optimisée. Cette exploration est souvent effectuée via des algorithmes génétiques [TODO :Lien(s) qui explique les principes] car ils correspondent à la problématique d'exploration d'un espace de solutions de grande dimension. Cette exploration est effectuée de manière non dirigée tout en fournissant un résultat exploitable.

Les réelles avancées dans le domaine sont récentes, les premiers articles concrets datent de 2016, et il est difficile de trouver des exemples dans un domaine concret, prouvant l'efficacité de *l'Automated Machine Learning*. Les créateurs de bibliothèque TPOT [2] ont rédigé deux papiers [5] [6] d'exemple d'applications dans des cas réels, sur la classification de cas de cancers de la prostate, de manière conventionnelle, et via l'approche *Automated Machine Learning*, et ont pu mettre en avant une amélioration des résultats. Google a récemment communiqué son intérêt pour le domaine, en annonçant l'ouverture d'un département sur la recherche de cette discipline [7]. Certains sites spécialisés [8] [9] décrivent ce domaine avec intérêt, mais en précisant que les résultats ne sont pas encore probants, et que, pour le moment, elle n'est pas applicable à toutes les problématiques.

Dans le cadre du projet, étant donné que les utilisateurs ne sont pas experts dans le domaine du *Machine Learning*, il est que les résultats soient meilleurs que les configurations des utilisateurs.

Si le travail abouti à une expérience, il est possible que celui-ci soit publié.

3.3 Technologies

3.3.1 TPOT

TPOT [2] est une bibliothèque *open-source* permettant l'optimisation de pipeline automatisée, alias *automated Machine Learning*. Elle se distingue des autres bibliothèques telles que Auto-WEKA [10] et Hyperopt [11] par le fait qu'il est capable non seulement de faire varier les modèles et le hyper-paramètres, mais qu'elle est aussi capable de sélectionner, construire ou d'effectuer du préprocessing sur les caractéristiques. TPOT dispose d'une communauté active, et le créateur *Randy Olson* répond très rapidement aux issues postées sur le *Github* de TPOT.

Cela se représente comme sur la figure figure 3.3.

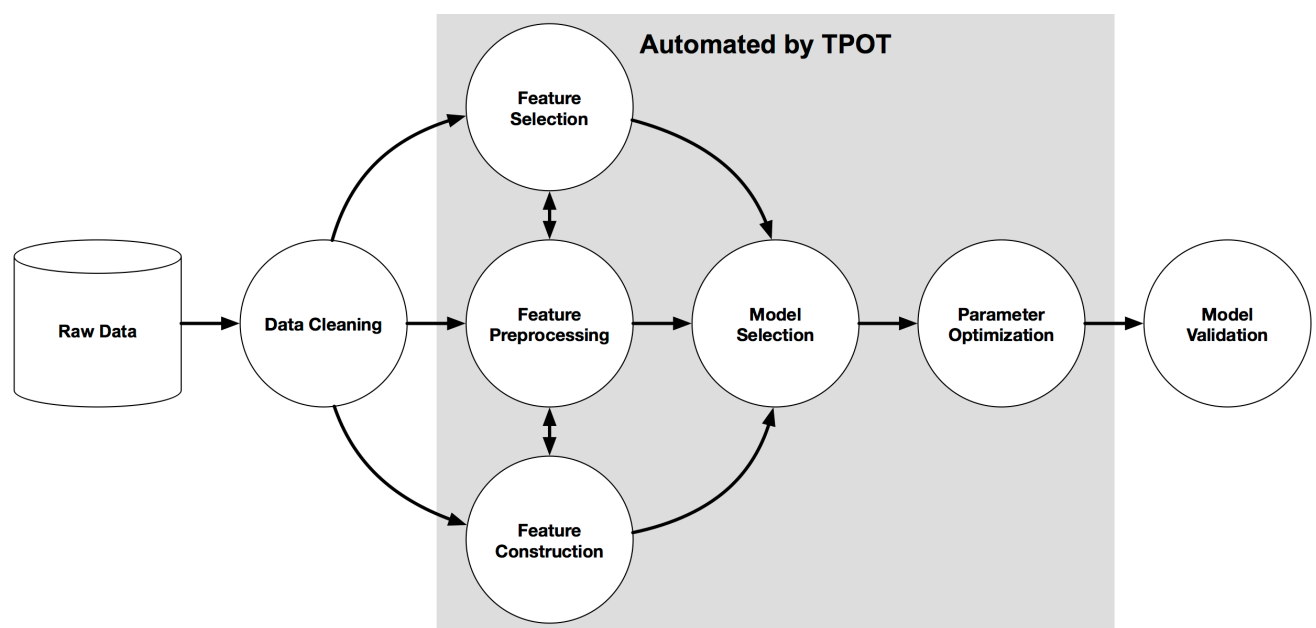


Fig. 3.3 : Exemple d'un pipeline de Machine Learning, avec les parties gérées automatiquement par TPOT. Illustration tirée de la documentation TPOT [2]

Les différentes étapes ont la même signification que présenté au point de *présentation des phases typiques de machine learning*, au dessous de la figure 3.2.

Cette bibliothèque est codée en Python, et se base sur les modèles de Scikit-learn [4], ce qui permet d'avoir des résultats exploitables directement via cette bibliothèque *Python*. Pour son implémentation, elle se base sur une représentation du pipeline sous forme d'arbre pour les pipeline (qui correspondent aux chromosomes dans la théorie de *Darwin*), et effectuer des mutations en croisant des parties de cet arbre, en coupant des branches, ou en créant de nouvelles.

3.3.2 Systèmes distribués

Historiquement, avant que le web ne vienne changer la donne, une application était localisé sur une machine unique, et son architecture se présentait comme sur la figure 3.4

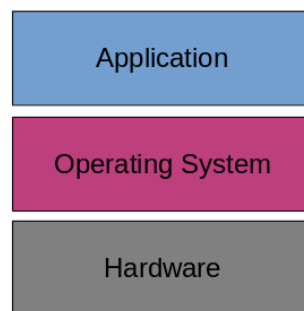


Fig. 3.4 : Architecture simple basée sur une application unique : crédits @ Groovytron [12]

Avec l'augmentation de la demande, la première approche pour augmenter la capacité de réponse a été de paralléliser plusieurs machines sur le réseau, et d'effectuer un balancement de charge entre les différentes instances, en fonction des moyens.

Les machines sont déployées en cluster (groupes de machines), et le *load-balancer* s'occupe de répartir les requêtes, comme présenté à la figure 3.5.

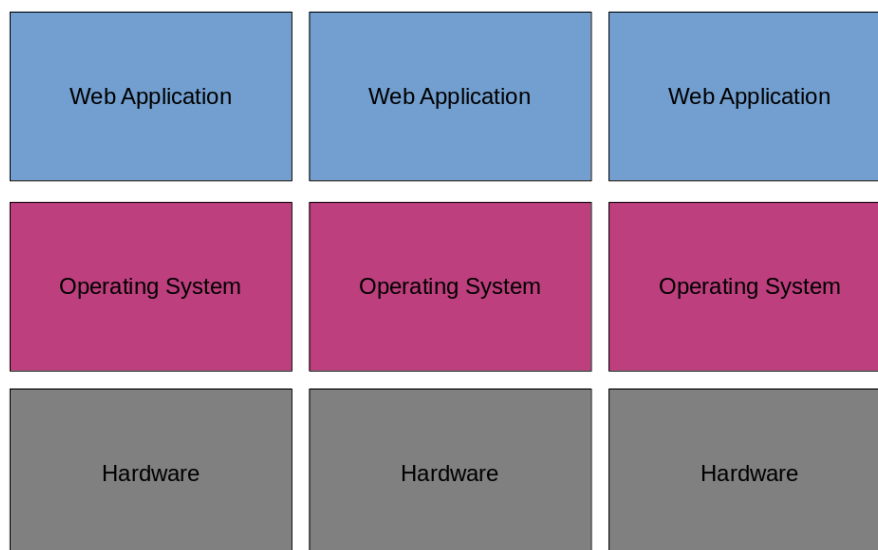


Fig. 3.5 : Architecture orientée haute disponibilité et « scalabilité » : crédits @ Groovytron

Avec la venue d'internet, l'utilisation des applications a changée, et elles ont été amenées à communiquer entre elles, afin de partager des données ou des services.

Dès lors, le découpage des applications s'est effectué par bloc, chaque application étant indépendante, mais fourni une interface comme point d'entrée pour communiquer, et s'appuie

généralement sur un format d'encodage haut-niveau (XML, JSON, ...). pour formuler des réponses aux autres applications. On a ainsi un découpage plus fin des fonctionnalités, mais ce découpage engendre un travail supplémentaire pour le programmeur.

Etant donné que les machines sont indépendantes, la gestion des ressources s'effectue pour chacune en local. Dans l'approche d'un système distribué, on cherche à pouvoir gérer le plus finement les ressources au niveau du cluster, et pas uniquement par un balanceur de charge.

La mise en place de systèmes d'exploitation distribués tels que *DC/OS* est un système qui se superpose au système d'exploitation de la machine, et qui fournit une gestion fine des ressources. la [figure 3.6](#) permet d'illustrer cette architecture.

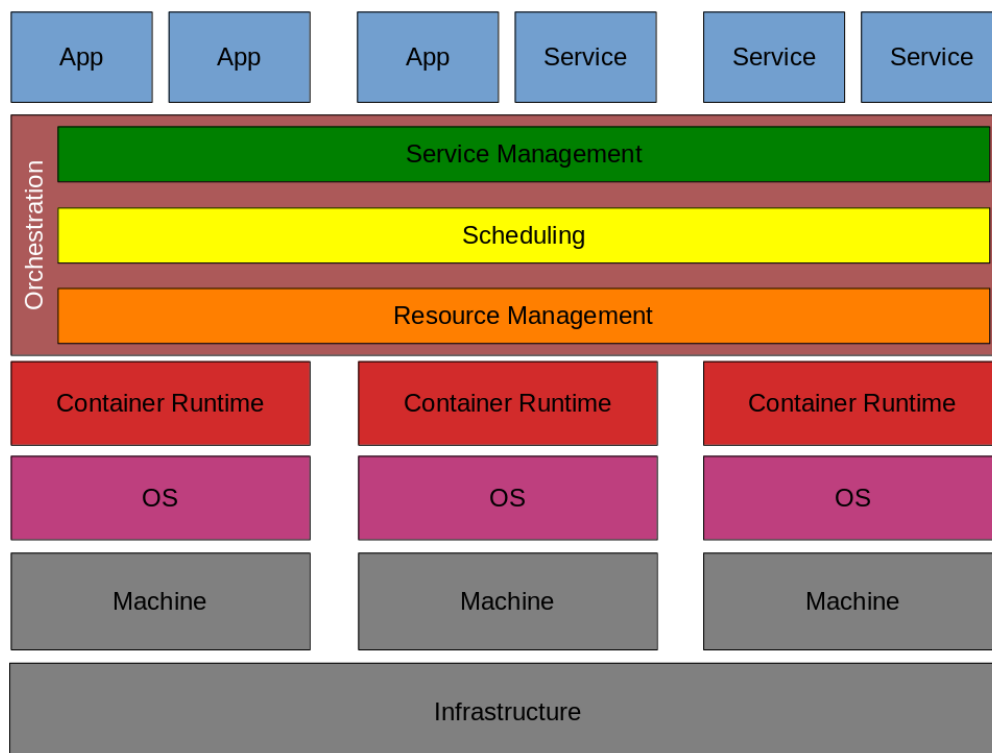


Fig. 3.6 : Architecture utilisant un outil d'orchestration de containers : crédits @ Groovytron

DC/OS est issu de la *Mesosphere*, un ensemble d'outils fournis par Apache qui répondent spécifiquement aux problématiques du cloud-computing. L'architecture du CHUV est basée sur les outils de la *Mesosphere*, mais n'utilise pas *DC/OS* au complet.

Les outils utilisés dans le cadre du projet sont décrits dans la suite du document.

3.3.3 Mesos

Élément central de l'architecture distribuée utilisée au CHUV, *Mesos* [13] est un noyau exécuté sur chaque machine du cluster, qui fournit une abstraction des ressources des machines du cluster. Il est ainsi possible de lancer une application en définissant la quantité de mémoire

vive, le nombre de processeurs, et l'espace disque à disposition, et Mesos s'occupe de gérer les ressources et la localisation de celles-ci, mais aussi de gérer le redémarrage de services en cas de pannes, et la mise à l'échelle d'un service.

Il permet de lancer des applications natives, mais aussi des containers Dockers, comme c'est le cas dans ce projet.

Le cluster est organisé sous la forme d'un noeud *Master*, et de noeuds *Slaves*. Le noeud *master* est responsable de recevoir les demandes d'instanciations de services, et il envoie les ordres aux noeuds *Slaves* approprié, selon les ressources disponibles. La communication entre le *Master* et les *Slaves* est effectué via *ZooKeeper*, qui est un système de stockage clé-valeurs dans un système de fichiers, ce qui permet de partager les configurations des différents acteurs de l'architecture.

Mesos sert donc de support pour l'instanciation de services sur notre architecture distribuée.

3.3.4 Marathon

Marathon est un logiciel développé par *Apache* dans le cadre de la *Mesosphere*. La Mesosphère est l'ensemble des qui sont utilisés dans le cadre de *DC/OS*, et qui sont officiellement soutenus par la fondation *Apache*. *Marathon* joue le rôle de surcouche à Mesos afin de simplifier le déploiement de **services longues durées**, c'est à dire qu'une définition de tâche adressée à *Marathon* concerne un certain nombre d'instances de ce service, et que si une instance vient à se stopper, *Marathon* va automatiquement relancer une instance de ce service.

Le logiciel fournit lui aussi une *API REST* [14].









3.3.5 Chronos

Chronos est un logiciel développé par la *communauté* Mesos. Cette communauté, contrairement à la *Mesosphere*, n'est pas officiellement soutenue par *Apache*, mais est constituée de gens ayant des intérêts pour des outils liés à la Mesosphère, et qui collaborent en suivant le développement des outils de la *Mesosphere*. La pérenité de ces outils ne sont donc pas garantis.

Chronos fait office de remplacement à *cron* de *Linux*, qui est un service permettant de planifier des commandes à effectuer à intervalles réguliers. *Chronos* permet d'effectuer le même travail sur un système distribué via *Mesos*.

Il s'oppose à *Marathon* dans son utilisation, car il permet de lancer une commande ou un container de manière spontanée, ou programmée, mais qu'il ne cherchera pas à garder en tout temps un certain nombre d'instances en cours d'exécution.

Il fournit une interface graphique [cf [figure 3.7](#)] permettant de programmer une nouvelle tâche planifiée, mais aussi une *API REST* permettant l'automatisation programmatique de création de tâches.

| + ADD JOB | | | | | |
|--|----------|---------|-------|---|---|
| SUCCESS | 4 | FAILURE | 0 | FRESH | 0 |
| RUNNING | 0 | QUEUED | 0 | IDLE | 4 |
| JOB | NEXT RUN | STATUS | STATE | ACTIONS | |
| python_mip_tpot_b00150ed_8868_49d5_891f_0991ef3dbc61 | OVERDUE | success | idle |  |  |
| python_mip_tpot_f6d46887_12a0_4a5b_9949_c886e1d94d2f | OVERDUE | success | idle |  |  |
| python_mip_tpot_16d99b1b_bc86_40a5_90cd_22c64d979524 | OVERDUE | success | idle |  |  |
| python_mip_tpot_dd5ec681_3b0a_4f39_8c94_d67b29ced6bc | OVERDUE | success | idle |  |  |

© 2016 Mesosphere, Inc.

Fig. 3.7 : Capture d'écran de l'interface graphique de Chronos

3.3.6 Docker

Docker est une solution *open-source* qui permet d'embarquer une application dans un container *Linux* qui peut être exécuté sur n'importe quelle machine.

Dans une deuxième mesure, il fournit des mécanismes pour rendre un container proche de la virtualisation, en permettant d'isoler les containers entre eux, mais tout en fonctionnant sur le même système hôte. Ceci a l'avantage par rapport à la virtualisation de ne pas embarquer le système d'exploitation pour chaque container virtuel, ce qui réduit la taille des images. En revanche, étant donné que le système d'exploitation est partagé, et malgré les mécanismes d'isolation entre container et hôte, il est très difficile d'arriver à un niveau de sécurité identique à celui des machines virtuelles, qui elles peuvent être sécurisées jusqu'aux niveau des instructions micro-processeurs.

Docker s'utilise généralement pour uniformiser les conditions de développement, car on peut dire qu'une image fonctionnant en *stand-alone* (c'est à dire sans interactions avec le système hôte, comme par exemple un montage de volume) doit fonctionner sur une autre instance Docker.

D'un point de vue haut niveau, un container est par défaut isolé de l'hôte au niveau :

- du réseau ;
- du système de fichier ;
- des paquets installés ;
- des services ;
- des utilisateurs.
- des processus (en partie) ;

Ce qui n'implique pas qu'il est impossible d'accéder à ces différentes instances de l'hôte, plus ou moins sciemment.

Docker se base sur un système d'image et d'héritage. Il est possible de créer son image personnalisée à partir d'une image minimale fournie par la communauté comme :

- BusyBox ;
- CentOS / Scientific Linux CERN (SLC) or Debian/Ubuntu or CentOS/RHEL/SLC/etc ;
- Debian / Ubuntu ;

Mais aussi from scratch, ou via une archive [15].

De plus, on peut hériter de n'importe quelle image et la redéfinir via sa propre surcouche. Les images dont on hérite ne sont pas modifiables. L'héritage est possible pour toute image publiée sur un dépôt d'image Docker (généralement Docker Hub, mais on peut en utiliser d'autres).

La spécialisation du comportement d'un container Docker s'effectue via un fichier de définition, le *Dockerfile*. Ce fichier est constitué de commandes [16] qui peuvent effectuer des actions pour construire l'image, dont les principales sont :

- Définir de quelle image on hérite ;
- Copier un fichier de l'hôte à l'intérieur du système de fichier interne ;
- Exécuter une commande bash ;
- Définir des variables d'environnements ;
- Définir quels ports on veut exposer à l'hôte.

Si on souhaite pouvoir choisir entre plusieurs commande, on peut définir des entrypoints, qui définissent en général un script que l'on peut exécuter suivant les paramètres d'appels du container.

Il y a deux méthodes d'interaction avec un container :

- `docker run`
- `docker exec`

La méthode *run* instancie le container. Il permet de définir des paramètres qui définiront des caractéristiques internes ou externes du container. Il est par exemple possible de définir le nom du container, un argument d'entrée (utilisable par l'entypoint) ou des variables d'environnement. Suivant l'implémentation du container, il est possible que celui-ci agisse comme un service, et se maintienne en vie, en attente de nouveaux événements, ou qu'il se termine dès que le travail interne soit terminé. Dans les deux cas, il se contente d'attendre que le travail interne(souvent implémenté par un script) renvoie un code d'erreur [17]

La méthode *exec* ne peut s'appeler que sur un container qui a déjà été instancié. Si le container est en cours d'exécution, il est possible d'envoyer une nouvelle commande au container. La plus classique est l'exécution d'un bash en mode interactif, via la commande : *docker exec -it containername bash*

qui permet d'exécuter une ligne de commande bash. Le paramètre *-it* permet justement de laisser la commande en mode interactif, ce qui permet de ne pas fermer l'exécution de la commande dès que celle-ci renvoie un code 0.

TODO :Finir cette section.

3.3.7 Scala

Ce travail est effectué au cœur du projet Woken du Human Brain Project. Ce projet contient le langage de programmation Scala [18]. Scala a été conçu à l'école polytechnique de Lausanne (EPFL) afin de proposer de lier des paradigmes de programmation différents et habituellement opposés, tels que la programmation fonctionnelle et la programmation orientée objet. Scala se base sur la JVM3, ce qui permet de bénéficier de l'abstraction de celle-ci en termes de plateforme d'exécution, ainsi que pour la gestion de la mémoire, notamment. Scala coopère ainsi de manière transparente avec Java, ce qui permet d'utiliser des bibliothèques non codées en Scala.

Cette section ne précise pas la syntaxique du langage, ni son utilisation.

3.3.8 AKKA

Akka [19] est un outil de développement et un environnement d'exécution libre et open-source qui a pour but de simplifier la mise en place d'applications distribuées et concurrentes basée sur la JVM. Il gère donc les langages de programmations Java et Scala, et est développé en Scala. Akka propose une résolution des problèmes de concurrence via un système d'acteurs.

Chaque acteur propose des fonctionnalités, et peut communiquer avec les autres en envoyant des messages. Lorsqu'un acteur reçoit un message, il le traite, effectue des actions et peut envoyer d'autres messages, instancier d'autres acteurs ou encore se stopper.

Chaque acteur est un client léger, qui possède son état et sa boîte aux lettres. Lorsqu'un acteur plante, il est réinstancié automatiquement, dans le même état qu'il était avant, et avec sa file de message, ce qui procure une haute disponibilité. De plus, lorsqu'un acteur plante, le parent est notifié, et il peut dès lors prendre des mesures. Les messages sont asynchrones, ce qui permet de ne pas avoir d'état bloquant en cas de latence réseau ou tout autre problème technique. Akka s'occupe de distribuer les acteurs sur le cluster, ce qui permet d'avoir un haut niveau d'abstraction pour le programmeur.

Cette section vise à décrire le cadre logiciel dans lequel le travail sera effectué, et à préciser les acteurs ainsi que leurs fonctions.

4.1 Woken

Woken [20] est un service, utilisable via une *API REST*, qui fournit la possibilité d'explorer les données (*data mining* en anglais) de la plateforme. Cette exploration de données peut être de différentes natures, comme ériger un graphe qui permet à l'utilisateur de visualiser les données, de demander une analyse statistique, d'effectuer une expérience de classification via un des algorithmes de classification fourni, ou encore une expérience de régression.

Chacune de ces explorations de données est effectuée sur un ensemble de données, qui est qualifié par les champs configurés dans la Fig. 1.3 par l'utilisateur de la plateforme.

Une expérience fournit des résultats au service demandeur, sous format PFA [21], qui est un format dont la syntaxe est basée sur *yaml*, mais dont la structure est destinée à décrire des pipeline pour le data-mining.

Des requêtes HTTP sont mises à disposition dans le répertoire `dev-debug/http` ou `dev-test/http` afin de permettre de se passer de l'interface graphique, et de simplifier le développement.

4.1.1 Place de Woken dans l'architecture

Woken étant un service, il est conçu pour être utilisé par d'autres services. La figure figure 4.1 présente une version simplifiée de l'architecture de la plateforme MIP.

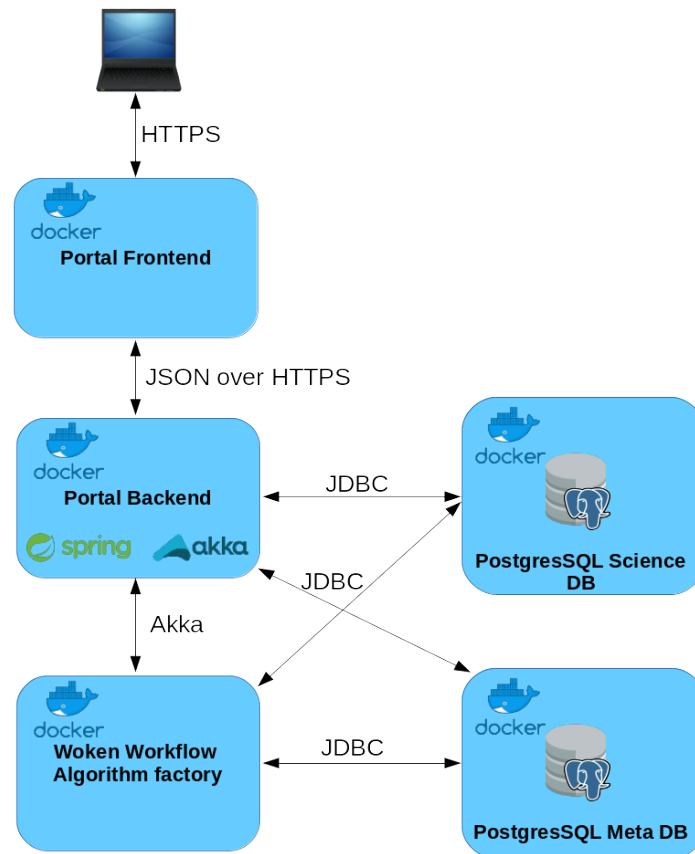


Fig. 4.1 : Architecture globale simplifiée de la plateforme MIP. Le **Portal Frontend** est le point d'accès pour l'utilisateur. Il peut consulter les données et effectuer des expériences depuis celle-ci. Le **Portal Backend** fournit les mécanismes de sécurité et d'accès aux bases de données, ainsi que la répartition des demandes à **Woken** si nécessaire. La base de données **Science-db** contient les données des patients, tandis que la base de données **Meta-db** contient les descriptions de chaque *feature* disponible dans la plateforme. Cette description permet de déterminer différentes informations pour la plateforme telles que le type de données (nominale ou continue) ou l'unité de mesure. Si il s'agit d'une demande nécessitant un algorithme, c'est **Woken**, l'*algorithm factory*, qui va s'occuper de traiter les demandes. **Woken** peut lui-aussi accéder aux bases de données afin d'appliquer ses algorithmes. Cette figure est une représentation simplifiée de l'architecture, qui ne contient pas tous les intervenants de la plateforme, mais uniquement ceux utilisés à cette échelle.

Lorsque l'utilisateur adresse une requête HTTP contenant une demande d'expérimentation, le backend envoie une demande de *mining* à *woken* via une requête POST de la forme :

```
POST localhost :8087/mining/job \
  variables := '["code": "cognitive_task2"]' \
  grouping := '[]' \
  covariables := '["code": "score_math_course1"]' \
  filters := '[]' \
  algorithm := '{"code": "knn", "name": "KNN", "parameters": []}'
```

Ce qui correspond aux paramétrage de l'expérience de l'utilisateur, comme présenté en [figure 1.3](#). *Woken* effectuera la requête, effectuera l'algorithme et retourne une réponse sous format *PFA*. Le format de réponse n'est pas important dans le cadre de ce projet. Il existe deux routes REST pour demander à *Woken* d'effectuer un travail :

- /mining/job
- /mining/experiment

Le /mining/job permet de lancer un seul algorithme à la fois, et permet pas de lancer des expériences utilisant la *cross-validation*, tandis que la route /mining/experiment permet la *cross-validation* et de lancer plusieurs algorithmes.

La requête pour une expérience via la route /mining/experiment se présente sous la forme :

```
POST localhost :8087/mining/experiment \
  variables := '["code": "cognitive_task2"]' \
  grouping := '[]' \
  covariables := '["code": "score_test1", {"code": "college_math"}]' \
  filters := '[]' \
  algorithms := '["code": "knn", "name": "knn", "parameters": []]' \
  validations := '["code": "kfold", "name": "kfold", "parameters": [{"code": "k",
↪ "value": "2"}]]'
```

4.1.2 Fonctionnement interne de Woken

Woken a la responsabilité d'appliquer des algorithmes suite à la demande via l'une des deux *routes REST* mises à disposition.

La [figure 4.2](#) présente les intervenants liés à *Woken* lors d'une demande d'algorithme.

Par rapport au code de *Woken*, le principal intervenant est le flux d'acteurs *Akka*, implémenté dans le script /src/main/scala/core/coordinator.scala. C'est celui-ci qui reçoit les expérimentations à effectuer. Celles-ci sont déterminées par un code d'algorithme, des *features* concernées, les variables cibles ainsi que le modèle et les hyperparamètres pour les expériences de *machine learning*.

Les acteurs *Akka* implémentés dans cette portion de code Scala héritent d'une méthodologie FSM (Finite State Machine), ce qui rend les acteurs capables de se comporter comme un automate à états finis. Les transitions entre ces états s'effectuent via des événements précis. Cette implémentation permet de mettre un acteur parent en attente des résultats des acteurs enfants, de manière élégante et sans attente active bloquante.

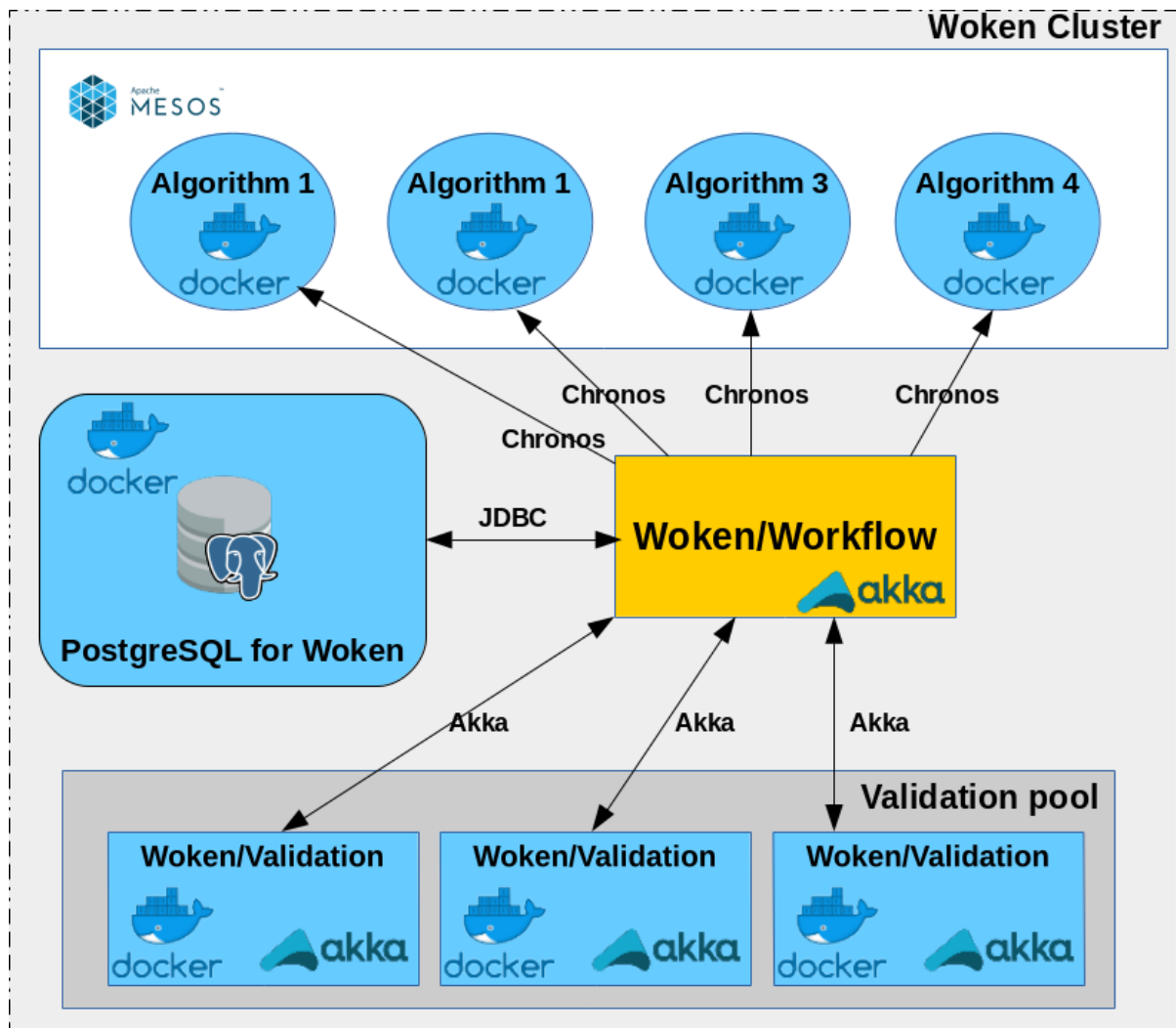


Fig. 4.2 : Architecture « interne » de woken. Les intervenants ici présents sont ceux qui sont directement utilisés par le service *Woken*. Le service **Woken** en lui-même est généralement contenu dans un container, mais il peut être en natif dans l'architecture, comme dans le cas de l'architecture dev-debug du projet. Il est responsable d'instancier des **algorithmes** contenus dans des containers *Docker*, via *Chronos*. Si l'expérience utilisateur demande une *cross-validation*, un **pool d'acteurs AKKA** s'occupant de cette tâche est instanciée au lancement de l'architecture, et sont prêts en tout temps à répondre à cette tâche. Ce choix a été effectué afin d'éviter de devoir instancier ces acteurs pour chaque demande, en prévision d'une forte charge sur la plateforme. Ces acteurs sont contenus dans un container *Docker*, ce qui permet de mettre à l'échelle en cas de besoin. Le résultat de chaque expérience est stocké dans la base de données **Woken-DB**, ce qui permet de récupérer le fichier de définition *PFA* afin de reconstruire l'expérience et de la vérifier, si besoin. Attention, les intervenants décrits ci-contre ne sont pas contenus dans le projet, mais bel et bien indépendants, et liés via la configuration *Docker-compose*.

Suivant la *route REST* en question, il existe deux flux possibles.

La route `/mining/job` se contente de lancer un *coordinatorActor*, qui est un acteur responsable de convertir un *Job* (case class scala) en JSON mis en forme selon le format d'entrée de Chronos, de lancer la requête à celui-ci, d'attendre les résultats dans la base de données *Woken-DB*, de mettre en forme le résultat et de le retourner au service demandeur.

La voie intéressante dans le cadre de ce projet est celle de `/mining/experiment`. Celle-ci a pour caractéristiques de pouvoir gérer plusieurs algorithmes pour une expérimentation, ainsi que de gérer la *cross-validation*.

Le flux de travail entre les acteurs peut être représenté comme montré sur la [figure 4.3](#)

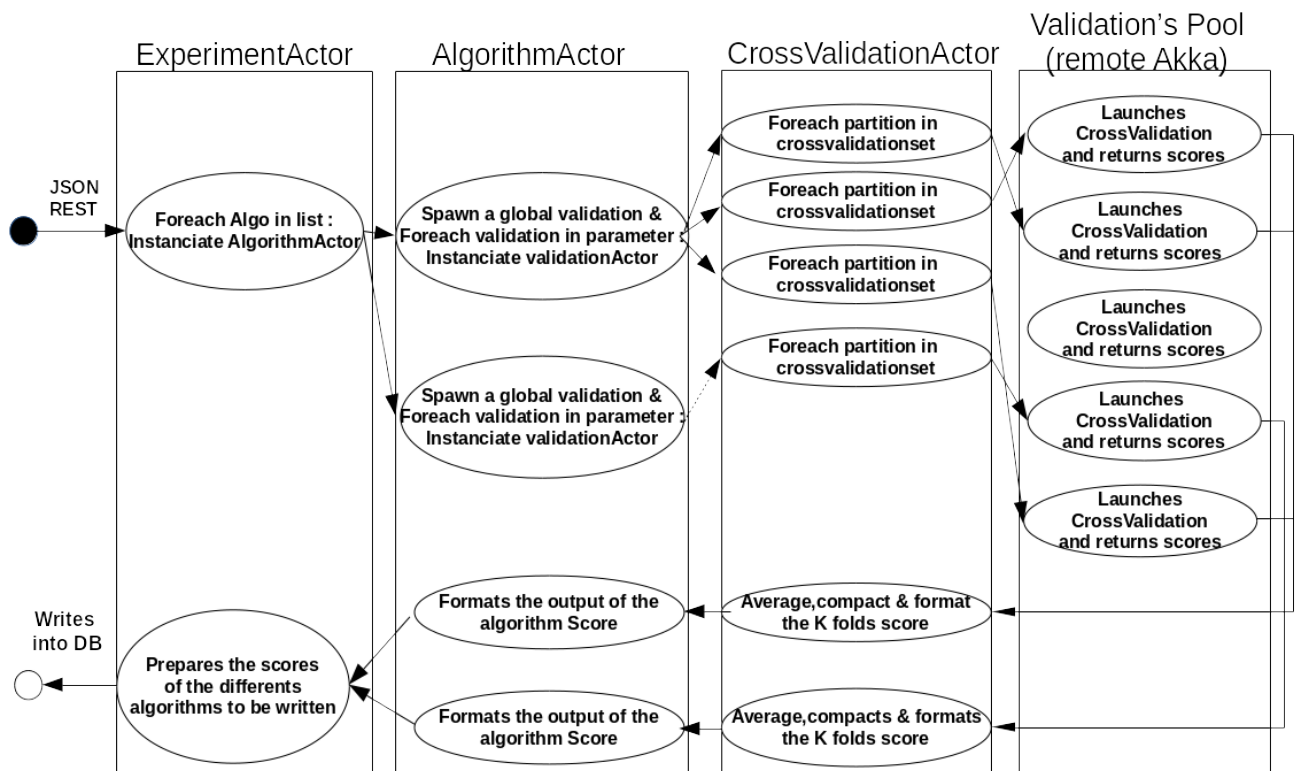


Fig. 4.3 : Schéma des acteurs Akka du script `coordinator.scala`. Les différents acteurs peuvent instancier d'autres acteurs dynamiquement, ce qui permet de répondre à n'importe quelle configuration d'expérimentation de l'utilisateur. Ce schéma correspond à l'implémentation avant le début du projet. Il existe des *coordinatorActor* qui ne sont pas documentés dans ce diagramme. Ils ont pour but d'envoyer un *Job* à Chronos, et d'attendre les résultats du container dans la base de données. Ceci induit qu'il n'y a pas de communication entre *Woken* et les containers. La cross-validation n'est effectuée que si l'algorithme est défini dans *Woken* comme prédictif. Les requêtes SQL sont envoyées aux configurations de Chronos (format JSON) sous la forme de variables d'environnement. Il est prévu de limiter l'accès aux bases de données à l'avenir, en passant le dataset aux containers, plutôt que de les laisser accéder directement aux bases de données.

Ce flux de travail comporte oblige deux problématiques de taille :

- Il est nécessaire d'attendre un résultat dans la base de données pour que les *localCoordinatorActor* détectent que le container a fourni un travail.

- Il est nécessaire de passer par le mécanisme de *cross-validation* pour définir un score à une expérience. Ceci impose aussi un format *PFA* strict.

4.1.3 Fonctionnement actuel des containers Docker

Actuellement, les containers utilisés par la plateforme Docker sont lancés via Chronos. A partir d'une définition d'expérience au format *JSON*, on instancie un objet de définition de cet algorithme en case classes Scala. Depuis ces définitions de classes, Woken sérialise en *JSON* correspondant au format attendu par Chronos, comme par exemple :

```
{
  "id": "chronos",
  "cpus": 1,
  "mem": 1024,
  "instances": 1,
  "container": {
    "type": "DOCKER",
    "docker": {
      "image": "mesosphere/chronos",
      "network": "BRIDGE",
      "portMappings": [
        {
          "containerPort": 4400,
          "hostPort": 0,
          "servicePort": 4400,
          "protocol": "tcp"
        }
      ]
    }
  },
  "volumes": []
}
```

Woken est actuellement capable d'instancier autant de containers que demandent les utilisateurs. Il s'occupe de générer des identifiants uniques comme *id* de tâche à Chronos, de récolter chacun des résultats dans la base de données et de les mettre en relation avec la bonne expérience.

Il n'est en revanche pas capable de communiquer avec un container. Une fois le fichier de configuration *JSON* envoyé via une requête POST HTTP, il ne peut qu'attendre les résultats dans la base de données.

Dans le cadre de notre nouveau flux, nous devons pouvoir attendre la fin du travail d'un container, récupérer son résultat, puis adresser une deuxième requête sur le container.

Cette fonctionnalité, que l'on peut qualifier de container « interactifs », doit faire l'objet de recherches. Docker est conçu pour être *stateless*. Quand un container meurt, si il n'a pas de *Volume* configuré, le container n'a pas de moyen d'enregistrer l'état dans lequel il était. Un *Volume* est un répertoire partagé entre le container et l'hôte. Si il existe des fichiers dans le dossier au moment du montage du *Volume*, le container y aura accès. Si un le container meurt,

le contenu du *Volume* reste.

4.1.4 Tests préliminaires avec TPOT

Des tests ont été effectués avec TPOT afin de déterminer son efficacité et son utilisabilité. Dans le cadre de du projet, le plus important était de pouvoir :

- Travailler avec le dataset du projet ;
- Pouvoir extraire le meilleure pipeline à la fin de l'optimisation ;
- Obtenir les scores ;
- Reconstruire le pipeline à partir via Scikit-learn.

Mais aussi, si possible :

- Déterminer les *features* construites ;
- Déterminer l'importance de chaque feature.
- Déterminer l'exploitabilité de l'export implémenté dans TPOT.

Des tests [22] ont été implémentés, et une issue [tpotissue] a été adressée sur le projet afin de vérifier les points délicats.

A la fin de ces tests, il s'avère que :

- Il est possible de récupérer le meilleur pipeline trouvé, avec les hyperparamètres du modèle, ainsi que son score ;
- Il est possible d'enregistrer un pipeline sous forme de texte, et de le ré-instancier en objet Scikit-learn utilisable pour des prédictions ;
- Tpot n'est pas en mesure de mettre à disposition la selection, la construction et la normalisation de features. Celles-ci sont données par le modèle ;
- L'export n'est pas utilisable dans notre contexte ;
- Il est possible de travailler via le dataset du projet.

A la fin de cette analyse, les attentes envers TPOT dans le cadre du projet sont atteintes.

4.1.5 Le cas de Marathon

Durant ce projet, la substitution de *Chronos* par *Marathon* a été explorée. La raison est que *Chronos* n'est pas directement lié à la *Mesosphere*, et que son développement n'est pas assuré sur le long terme. Une fois intégré dans l'architecture, il s'est avéré que *Marathon* ne répond qu'à la problématique des services de longues durées.

Metronome [23] est destiné à être le remplaçant de *Marathon*, mais il n'est actuellement pas assez abouti pour être incorporé à l'architecture.

Cette section ne décrit que les choix de conceptions qui ont été implémentés.

A partir de l'analyse effectuée au chapitre analyse, il est possible de concevoir la nouvelle architecture pour résoudre les problématiques connues qui sont :

- Mettre en place un flux qui se passe de l'attente des résultats des containers dans la base de données ;
- Mettre en place un container qui accepte plusieurs points d'entrées ;
- Se passer de la mise en forme de PFA imposée dans le flux actuel.

5.1 Système de containers interactifs

5.2 Modification globale du workflow Woken

Sans parler directement de la modifications des acteurs *Akka*, il est intéressant de présenter une vue d'ensemble des intervenants dans la problématique du projet, et de définir les rôles qu'ils remplissent dans cette nouvelle conception.

La [figure 5.1](#) présente une représentation du flux imaginé.

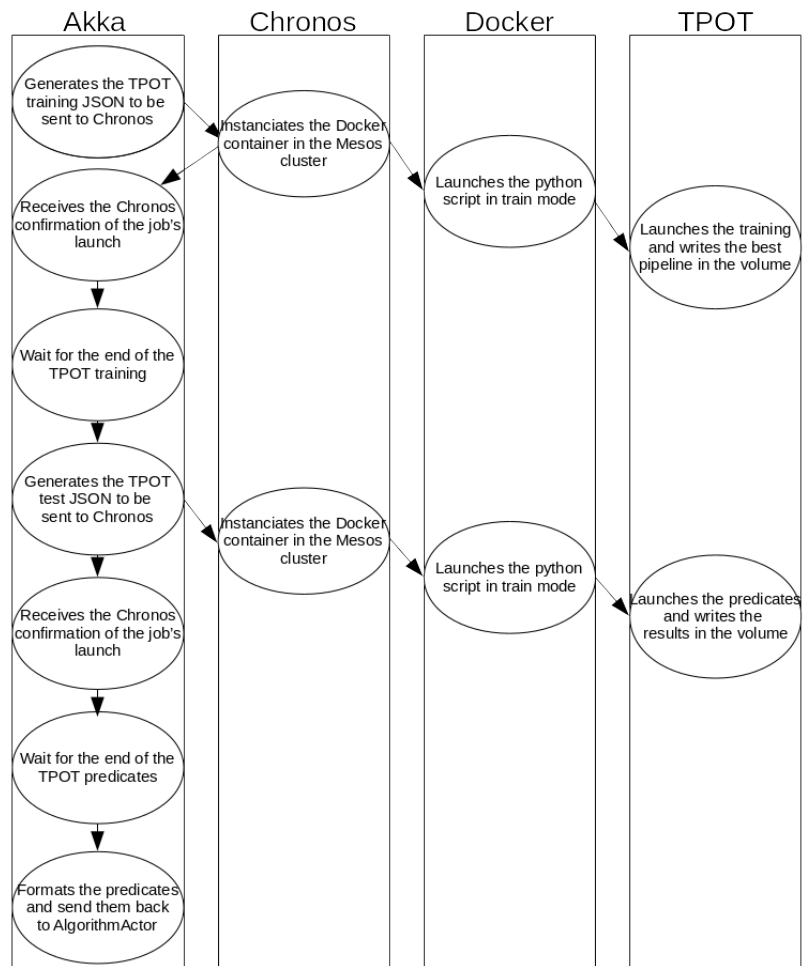


Fig. 5.1 : Schéma représentant les interactions des intervenants pour la conception retenue. Les titres de colonnes définissent la technologie responsable des tâches qui sont dans la colonne. Pour le cas d'Akka, cette représentation ne correspond pas à un diagramme d'acteur. Ce nouveau flux utilise les *volumes* de Docker afin de faire persister les résultats d'entraînements et de prédicats.

5.3 Nouveau diagramme d'acteurs imaginé, et comment on coupe le workflow actuel

6.1 Création d'un container interactif

6.1.1 Problème initial

6.1.2 Présentation des solutions au problème

6.1.3 Choix effectué

6.2 Modification du workflow Woken

6.2.1 Ajout du nouveau container dans la configuration

6.3 Intégration de TPOT

6.3.1 A déterminer, mais je suppose : Les contraintes posées par la bibliothèque, les choix qui ont du être effectués.

6.4 Eventuellement, si plus de travail a été effectué, présentation de celui-ci.

Validation (Expérience)

6.1 Présentation de l'expérience 6.1.1 pourquoi 6.1.2 comment 6.1.3 les conditions de tests
6.2 Résultats de l'expérience 6.3 Discussion des résultats

8.1 Etat des lieux au moment du rendu

- Atteintes des objectifs
 - Le contexte du mandant a-t-il été compris ?
 - L'API se superposant à Marathon fonctionne-t-elle ?
 - Un format de métadonnées a-t-il été spécifié ? Existe-t-il un moyen de vérifier que telle ou telle image Docker respecte ce format ?
 - Un démonstrateur a-t-il été développé ?
- Améliorations possibles

8.2 Perspectives et améliorations

8.3 Bilan personnel (Présenter ce qui apporte quelque chose)

Annexes, références et Table des illustrations.

TODO :Annexes : - CdC - Journal de travail - TPOT papers - FULL Rest API for TPOT

Bibliographie

- [1] Philippe Beraud - Microsoft France. Un peu de théorie pour l'apprentissage non-supervisé). https://blogs.msdn.microsoft.com/big_data_france/2014/06/06/un-peu-de-theorie-pour-lapprentissage-non-supervis/, June 2014.
- [2] Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore. *Applications of Evolutionary Computation : 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*. Springer International Publishing, 2016. ISBN 978-3-319-31204-0. URL : http://dx.doi.org/10.1007/978-3-319-31204-0_9, doi :10.1007/978-3-319-31204-0_9¹.
- [3] Scikit-Learn's Documentation. Tuning the hyper-parameters of an estimator. http://scikit-learn.org/stable/modules/grid_search.html, July 2017.
- [4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.
- [5] Randal S Olson, Nathan Bartley, Ryan J Urbanowicz, and Jason H Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, 485–492. ACM, 2016.
- [6] Randal S Olson and Jason H Moore. Tpot : a tree-based pipeline optimization tool for automating machine learning. In *Workshop on Automatic Machine Learning*, 66–74. 2016.
- [7] Google developers. Google i/o keynote (google i/o “17). <https://www.youtube.com/watch?v=Y2VF8tmLFHw>, May 2017.
- [8] Matthew Mayo. The current state of automated machine learning). <http://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html>, Jan 2017.
- [9] Hamel Husain. Automated machine learning — a paradigm shift that accelerates data scientist productivity @ airbnb. <https://medium.com/airbnb-engineering/automated->

1. https://doi.org/10.1007/978-3-319-31204-0_9

machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8, July 2016.

- [10] auto-WEKA. Automatic model selection and hyperparameter optimization in weka. <http://www.cs.ubc.ca/labs/beta/Projects/autoweika/>, July 2017.
- [11] hyperopt. Hyperopt - distributed asynchronous hyperparameter optimization in python. <http://hyperopt.github.io/hyperopt>, July 2017.
- [12] Julien M’Poy / Groovytron. Maracker’s repository : api aiming to make human brain project’s medical informatics platform’s developed apps deployment on mesos marathon easier. <https://github.com/groovytron/maracker>, July 2017.
- [13] Apache Software Foundation. Apache mesos. <http://mesos.apache.org>, July 2017.
- [14] Apache Software Foundation. Marathon rest api. <https://mesosphere.github.io/marathon/docs/rest-api.html>, July 2017.
- [15] Docker Inc. Create a base image - docker documentation. <https://docs.docker.com/engine/userguide/eng-image/baseimages/#create-a-full-image-using-tar>, July 2017.
- [16] Docker Inc. Dockerfile reference - docker documentation. <https://docs.docker.com/engine/reference/builder/>, July 2017.
- [17] sallyom. Change “docker run” exit codes to distinguish docker/contained errors #14012. <https://github.com/moby/moby/pull/14012>, December 2015.
- [18] Switzerland Lausanne (EPFL) Lausanne. The scala programming language. <https://www.scala-lang.org/>, July 2017.
- [19] AKKA Lightbend Inc. Akka official website. <http://akka.io>, July 2017.
- [20] axelroy. Woken - an orchestration platform for docker containers running data mining algorithms - forked by axelroy on github. <https://github.com/axelroy/woken/tree/AutoML>, July 2017.
- [21] DMG - Data Mining Group. Pfa - portable format for analytics). <http://dmg.org/pfa/index.html>, July 2017.
- [22] Axel Roy. Tests with the tpot library github page. https://github.com/axelroy/TPOT_Tests, July 2017.
- [23] DCOS/Metronome - Apache. Apache mesos framework for scheduled jobs - github page. <https://github.com/dcos/metronome>, July 2017.
- [24] Harbur Cloud Solutions S.L. Captain - convert your git workflow to docker containers). <https://github.com/harbur/captain>, July 2017.
- [25] Axel Roy. Visualize constructed features and get best pipeline found. <https://github.com/rhiever/tpot/issues/459>, May 2017.