Projet Cassiopée
Plateforme pédagogique et de prototypage dédiée à l'IA et à la Data Science

Robert Voyer * Thierno Tounkara †

24 novembre 2021

Table des matières

*Institut Mines Télécom - Lab. LASCO Paris Descartes.
†Institut Mines Télécom - IS Lab.

1 Introduction

Le phénomène Intelligence Artificielle

L'« *Intelligence Artificielle* » (IA en abrégé) est partout, pour le meilleur comme pour le pire, suscitant autant d'espoirs que de craintes.

Tous les secteurs sont concernés, et les perspectives économiques considérables : l'IA pourrait contribuer à hauteur de 15 700 Mds de dollars à l'économie mondiale en 2030[1].

Longtemps considérée comme prometteuse, mais en réalité peu dynamique depuis les années 1960-1970, l'IA est désormais en plein essor. Avec le *deep learning*, la machine peut non seulement réagir de façon programmée à certaines situations mais aussi réagir aux résultats de ses propres décisions, les garder en mémoire, les analyser et réélaborer en retour ses propres modes de fonctionnement.

La machine est désormais capable de tirer des leçons de ses comportements passés, d'améliorer elle-même la qualité de ses réponses et, pour être toujours plus efficace, d'optimiser son propre fonctionnement ; on parle d'*apprentissage profond*.

Cette révolution a été permise par la combinaison de trois facteurs :

— L'apparition et la démocratisation de matériels informatiques de plus en plus puissants, principalement les « GPUs (Graphical Processing Units) capables de plus de mille milliards d'opérations par seconde ».

— La création des réseaux neuronaux artificiels (ensemble d'algorithmes inspirés du fonctionnement des neurones).

— La disponibilité de masses de données. L'accumulation de données permet à ces programmes de progresser continuellement.

Ainsi, la reconnaissance vocale, la traduction automatique et la reconnaissance d'image progressent à très grande vitesse.

Pénurie de spécialistes en IA

Tous les experts (IDC, Gardner Group, rapport Villani, libres blancs, ...), prévoient une pénurie importante de spécialistes en IA et Data Science d'ici peu (2020). Aussi, la demande en formations, conférences et conseils dans ces domaines et pour tous les publics, ou presque, ne cesse d'augmenter. Les écoles d'ingénieur ou de management, les universités ainsi que les organismes de formation ont d'ores et déjà intégré dans leurs offres des cursus IA, et plus spécifiquement *Machine Learning* et *Deep Learning*. Notre proposition de plateforme pédagogique a donc plusieurs objectifs :

— Participer au rayonnement de l'école dans le domaine de l'IA.

— Proposer des cours interactifs dédiés à l'IA et à la Data Science.

— Proposer une valeur ajoutée dans la transmission du savoir.

2 Vidéos - SPOC - MOOC

Face à la demande sans cesse croissante, le nombre de propositions de cours et de formations dédiés à l'Intelligence Artificielle ne fait qu'augmenter. Il faut cependant noter que la qualité est inégale et peu propose de véritable plateforme. Les cours et formations se présentent sous plusieurs formes :

— Vidéo

— Contenu Web

— SPOC - MOOC

Vidéos Youtube

Beaucoup adopte cette forme (souvent comme appât pour des formations payantes) dont les plus connus sont :

— *Edureka!*

— *Intellipaat*

— *Simplilearn*

Il faut noter que les supports (de type power point) sont souvent très pédagogiques et agréables visuellement.

On peut y inclure aussi tous les passionnés qui créent des chaines youtube spécifiques. À noter que pour certains les supports utilisés sont quasi inexistants et que pour d'autres, bien au contraire, ce sont de véritables prouesses graphiques. On peut citer :

[1] Soit une hausse respective de 26 %, de 14,5 % et de 9,9 % des PIB de la Chine, des États-Unis et de l'Europe du Nord.

— *3Blue1Brown*

— *Serrano.Academy*

Les vidéos Youtube, même très bien construites, ne sont pas des plateformes à proprement parler puisqu'ici il ne s'agit que de dérouler un enregistrement explicatif. C'est plutôt un type de ressource à intégrer dans une plateforme.

Contenu Web

On retrouve les mêmes acteurs qui proposent des vidéos youtube. Ces contenus sont en quelque sorte une version explicative souvent très pédagogique et belle graphiquement des présentations youtube. Les contenus se résument à des présentations web classiques.

SPOC - MOOC

Pour simplifier ce sont des cours contenant des vidéos (au même titre que précédemment) complétés de ressources (supports de cours, exercices, codes, données, etc.). Ce ne sont pas non plus des plateformes, mais plutôt des assemblages de plusieurs types de ressources indépendantes (vidéos, texte, codes).

Les acteurs proposant ce type de cours sont légions :

— Stanford - MIT - IMT - Universités et écoles - ...

— Udemy - Edureka ! - Intellipaat - Simplilearn - ...

— Microsoft- IBM - Google - ...

3 Plateformes d'enseignement

Une plateforme d'enseignement est une architecture qui intègre mais surtout orchestre plusieurs types de ressources lesquelles peuvent éventuellement interagir. Les plateformes se distinguent suivant l'éventail des ressources prises en compte ainsi que le niveau d'intégration. Ainsi, en partant du niveau d'intégration le plus bas :

— **Code interactif**. C'est déjà le cas pour tous ceux qui proposent des cours qui reposent sur l'interpréteur *Jupyter Notebook*. Le cours est ponctué de contenu python exécutable depuis un environnement notebook intégré. On a ici, une interaction qui permet à l'apprenant de mettre en pratique les notions qui viennent juste d'être délivrées. Les acteurs sont nombreux :

  — *kaggle*

  — *Colab*

— **Code + conseils et solution**. Certains environnement proposent de plus, la possibilité de délivrer directement à l'apprenant des conseils et la solution de l'exercice en cours.

— **Code + solution + simulation**. C'est le cas de la plateforme pédagogique de Google : *Cours d'initiation au Machine Learning*. Ce cours intègre les ressources suivantes :

  — vidéos

  — supports

  — exercices interactifs depuis *colaboratory*

  — solutions et conseils sur les exercices réalisés

  — configuration et évaluation de réseaux de neurones avec l'intégration et l'adapation du Playground

4 Plateformes de prototypage

On pense ici aux environnements de développement et de prototypage rapides comme *Amazon SageMaker* dont voici un extrait de présentation :

> Le développement traditionnel de machine learning est un processus complexe, onéreux et itératif, qui peut être encore compliqué par l'absence d'outils intégrés qui couvrent l'ensemble du flux de travail. Il est donc nécessaire de relier les outils et les flux de travail, ce qui prend du temps et génère des erreurs. SageMaker apporte une solution à ce problème en fournissant tous les composants utilisés pour le machine learning dans une unique boîte à outils. Ainsi, les modèles arrivent en phase de production plus rapidement, avec moins d'effort et à des coûts réduits.

5 Objectifs et résultats du projet 2020

Les objectifs du projet Cassiopée de l'année dernière étaient les suivants :

> L'objectif principal du projet est de concevoir *une plateforme pédagogique dédiée à l'enseignement de l'Intelligence Artificielle (IA) et de la Data Science*. Cette plateforme doit être le support pour des travaux pratiques encadrés et doit offrir la possibilité aux étudiants de faire de l'auto-apprentissage. A titre d'exemple, les étudiants pourront travailler en équipes et utiliser des objets connectés mis à disposition pour collecter des données, les transmettre via un opérateur d'objets connectés (SigFox par exemple) à la plateforme, les intégrer à un outil d'analyse de données, les analyser et en déduire une proposition de valeur. Au-delà d'un simple exercice pratique, les collectes et analyses seront réalisées dans le cadre de cas d'actualité, et

pourront aboutir à des interprétations et recommandations à forte valeur ajoutée pour les entreprises et la société. La plate-forme servira ainsi de support d'apprentissage technique et pragmatique des étudiants et renforcera les liens de l'école avec les acteurs de son écosystème.

Ce dispositif pédagogique sera mis à disposition des majeures IMT-BS, dans le cadre d'un module « Artificial Intelligence and Data Science » . Du côté des enseignants-chercheurs, cette plateforme constituera, à moyen terme, une opportunité pour accéder à des données à travers les cas d'usage réalisés par les étudiants et à des outils d'analyse.

Les résultats obtenus par l'équipe 2020 sont deux cas d'usage populaires du deep learning et qui ont été déployés chacun sur une plate-forme différente :

— *Vision numérique* : reconnaissance d'émotions sur un visage.
Plate-forme : ***Teachable Machine***

— *Traitement du langage naturel* : identification du sentiment (positif ou négatif) exprimé dans un tweet.
Plate-forme : ***Spacy***

Bien qu'intéressants et illustratifs, les résultats obtenus à l'issue du projet sont très limités puisqu'ils ne concernent que deux modèles de deep learning. Aussi, dans la continuité, le projet proposé cette année est plus ambitieux et résolument tourné vers l'évaluation et la mise en œuvre de plate-formes de prototypage dédiées à la *Data Science*. Pour rappel, un projet de Data Science implique de prendre en compte les étapes suivantes :

1. Accès aux sources de données.
2. Préparation des données.
3. Définition du ou des modèles d'apprentissage.
4. Évaluation du ou des modèles d'apprentissage.
5. Interprétation et visualisation des résultats.
6. Mise en production.

6 Objectifs du projet 2021

6.1 Évaluation des plate-formes

Il existe plusieurs plate-formes de prototypage[1] qui offrent des services plus ou moins développés. Le premier objectif du projet est d'établir une évaluation, points forts et points faibles[2], de plusieurs plate-formes[3] dont par exemple :

— *Orange3*
Orange est un environnement dédié à l'apprentissage automatique, à l'exploration et à la visualisation des données. Il est écrit en Python et permet facilement l'intégration de scripts dans ce langage[4].
Site officiel

— *DataRobot*
La plateforme professionnelle d'IA de DataRobot accélère et démocratise la data science en automatisant le parcours de bout en bout, des données jusqu'à leur transformation en valeur ajoutée[5].
Site officiel

— *Amazon SageMaker*
Amazon SageMaker est un service entièrement géré permettant aux développeurs et aux scientifiques des données de créer, de former et de déployer rapidement et facilement des modèles de machine learning. SageMaker facilite chaque étape du processus de machine learning afin de rendre plus aisé le développement de modèles de haute qualité.
Site officiel

— *Dataiku*
En tant que plateforme IA, Dataiku permet d'exploiter des ensembles de données à partir d'une architecture fondée sur l'apprentissage automatique.
Site officiel

— *RapidMiner Studio*
Rapid Miner est l'un des meilleurs systèmes d'analyse prédictive. Il est écrit en Java. Il fournit un environnement intégré pour l'apprentissage en profondeur, l'exploration de texte, l'apprentissage automatique et l'analyse prédictive.
Site officiel

— *Alteryx*

[1] *10 Best Data Mining Tools in 2021* 2021 ; *Top 15 Best Free Data Mining Tools : The Most Comprehensive List* 2021 ; *The 16 Best Data Science and Machine Learning Platforms for 2021* 2021 ; *Compare DataRobot vs Dataiku DSS* 2021.

[2] Le prix parfois très élevé devient le point faible essentiel et insurmontable (KNIME). Ce peut être aussi les fortes limitations entre la version d'essai gratuite et la version payante (RapidMiner).

[3] Les plate-formes indiquées ici figurent parmi les plus utilisées.

[4] *Data Mining Fruitful and Fun* 2021.

[5] *Plateforme professionnelle d'IA de DataRobot* 2021.

Alteryx[1] est un outil d'analyse et de traitement automatique de données, qui peut être employé par les entreprises et leurs collaborateurs spécialistes de la data, afin d'en faciliter la lecture et le traitement, puis l'établissement de process et de workflows automatisés à partir de ces dernières.
Site officiel

— *KNIME*
KNIME est une plate-forme open source pour l'exploration des données et l'apprentissage automatique. Son interface intuitive permet de créer de bout en bout des workflows Data Science, de la modélisation à la production. Différents composants pré-construits permettent une modélisation rapide sans entrer une seule ligne de code.
Site officiel

— *WEKA*
Également connu sous le nom de Waikato Environment, WEKA est un outil dédié à l'apprentissage automatique développé. Il est particulièrement adapté à l'analyse de données et à la modélisation prédictive. Il contient des algorithmes et des outils de visualisation dédiés à l'apprentissage automatique.
Site officiel

Nous conseillons de porter votre attention sur les trois environnements suivants :

— *Orange3*

— *DataRobot*

— *RapidMiner*

Cependant, en fonction de votre évaluation, le choix de l'outil à analyser en détail pourra être différent de l'un des trois que nous avons retenus.

Nous avons plus particulièrement étudié *Orange3* que nous présentons rapidement dans le paragraphe suivant.

6.2 Plate-forme Orange3

Mettre en place des ressources pédagogiques (basées, entre autres, sur nos cours) associées à une plate-forme restent l'objectif principal du projet. Nous avons pris connaissance d'une plate-forme de simulation, d'enseignement et de prototypage tout à fait adaptée à ces besoins : la plate-forme : *Orange3*.

Orange3 couvre quasiment tous les aspects du développement Data Science - Machine Learning. Elle est très visuelle et interactive comme l'illustre la figure 1.

[1]Point faible : pas de version Mac OS

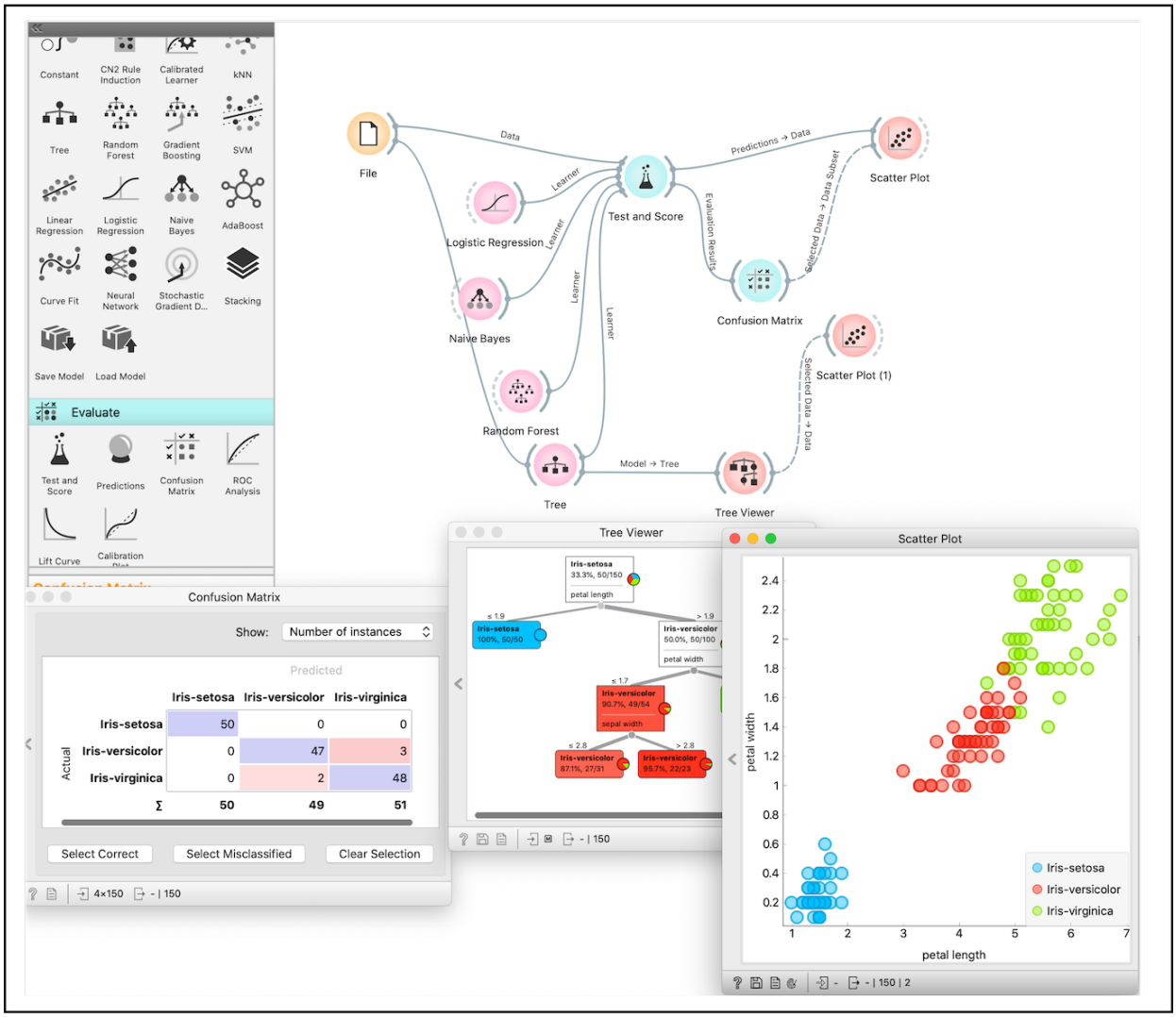

FIGURE 1 – Exemple Orange3

Orange3 est un environnement de développement de solutions de Data Science sans aucun codage (ou presque). On parle de plate-forme « *No Coding* ».

Les intérêts de la plate-forme « no coding » *Orange3* sont multiples et concernent les :

— *Étudiants* pour l'apprentissage du « ***Machine Learning*** ».

— *Chercheurs* pour utiliser dans leurs recherches les dernières techniques de « ***Machine Learning*** ».

— *Industriels* qui désirent mettre en œuvre des outils de « ***Machine Learning*** » pour résoudre des problèmes anciens afin d'en tirer de nouvelles opportunités à partir d'un vaste ensemble de données qu'ils détiennent déjà.

Orange3 permet très rapidement d'analyser un sous-ensemble de données (ou l'ensemble complet) pour en déduire des prédictions face aux problèmes qui se présentent. Même un programmeur python ou R peut bénéficier d'*orange3* pour expliquer graphiquement et par simulation ses choix algorithmiques.

Orange3 couvre toutes les étapes de mise en place d'un projet Data Science :

1. Récupération des données depuis un très large éventail de sources différentes.
2. Analyse et préparation sophistiquées des données.
3. Visualisation interactive des données et résultats obtenus.
4. Création de plusieurs modèles d'apprentissage.
5. Comparaison et test des modèles.
6. etc.

Pour terminer, *Orange3* propose aussi des modules interactifs spécialement conçus pour l'enseignement. Ce n'est pas anecdotique, ces modules sont très utiles en phase d'apprentissage car ils permettent de visualiser graphiquement le fonctionnement d'un principe (descente de gradient par exemple) ou d'un modèle de Machine Learning (K-Means, régression polynomiale par exemple).

7 Livrable du projet

Phases	Objectifs & Livrables
1. Évaluation des plateformes	Faire une évaluation des principales plateformes Les évaluations portent sur les mêmes études de cas Sélectionner la plus adaptée
2. Mise en œuvre de la plateforme *choisie*	Faire une évaluation précise de la plateforme permettant de supporter des solutions d'IA et de Data Science
3. Conception et tests de cas d'usage	Définir les objectifs d'apprentissage Concevoir les cas d'usage Réaliser les cas d'usage Tester les cas d'usage
4. Réalisation d'un pilote avec une majeure en apprentissage	Tester le dispositif pédagogique dans le cadre d'un cours d'IA avec des étudiants en M2 (cours réalisé par des enseignants-chercheurs)
5. Évaluation de la plateforme d'apprentissage sur la base du pilote	Collecter les retours d'expériences de la part des formateurs (grille d'entretien) Évaluer la perception des étudiants sur l'expérience d'apprentissage avec la plateforme (questionnaire en ligne) Évaluer qualitativement le niveau d'acquisition des compétences des étudiants avec la plateforme (questionnaire en ligne)

L'évaluation des plate-formes doit prendre en compte au moins les points suivants :

— Les mêmes études de cas doivent être utilisées pour chaque plate-forme.

— Qualité des outils d'assistance proposés sur les différentes étapes de mise en place d'un projet Data Science.

— Facilité de mise en œuvre des différents outils.

— Complexité conceptuel des outils.

— etc.

S'agissant des cas d'usage, il est surtout demandé de reprendre les exercices et études de cas que nous avons *déjà réalisés* dans le cadre de nos enseignement de l'« *Intelligence Artificielle et de la Data Science* ». Les exercices et études de cas sont actuellement définis dans des *notebooks Python* sur la plateforme *Colaboratory* de Google.

8 Ressources

— Un accompagnement sera prévu pour le binôme /trinôme travaillant sur ce projet

— Du matériel pourra être acheté pour mettre en œuvre certains cas d'usage (solutions industrielles complexes) très consommateurs en volume (plusieurs centaines de milliers de données) et en ressources machine.

9 Que vous apportera ce projet ?

Ce projet constitue une opportunité pour :

— Monter en compétences sur la mise en œuvre de l'IA et de la Data Science autour de cas d'usage réels.

— Comprendre les principes de fonctionnement de la quasi-totalité des modèles de Machine Learning supervisés et non supervisés :

  — Modèles supervisés.
  Régressions numériques et classifications : régression logistique, SVM, Arbres de décision, Random Forest, Gradient Boost, AdaBoost, Naive Bayes, etc.

  — Modèles non supervisés.
  PCA, K-Means, Hierarchical Clustering, Louvain Clustering, etc.

— Mais aussi comprendre et prototyper les phases de mise en place d'un projet de Data Science dont l'étape essentielle de préparation des données aussi appelée « *Wrangling* ».

  — Analyse des données aberrantes, des données manquantes.

  — Suppression des données inutiles.

  — Restructuration des données catégoriques.

  — Normalisation des données.

  — etc.

— Développer des compétences d'ingénierie pédagogique qui vous seront utiles dans votre future carrière

10 Client du projet

Ce projet Cassiopée est porté par une équipe deux enseignants-chercheurs du Département TIM, avec des expertises complémentaires décrites ci-dessous.

— **Robert VOYER** - Département TIM
Maître de conférences à Institut Mines-Télécom Business School, ses centres d'intérêt en recherche et enseignement portent sur la théorie des langages à objets, les modèles de représentation des connaissances et de leur intégration dans les systèmes d'information, les architectures d'entreprises J2EE, l'intelligence artificielle et les problématiques éthiques en entreprise.

— **Thierno TOUNKARA**
Docteur en Informatique de l'Université Paris IX (Dauphine), Thierno Tounkara est Maître de Conférences à Institut Mines-Télécom Business School. Il mène ses recherches à la jonction de plusieurs domaines : Ingénierie et Management des Connaissances, Système d'Information et Aide à la Décision. Il a également une forte expertise en ingénierie et en gestion de projet pédagogique. Il a déjà obtenu plusieurs financements du CFA pour des projets pédagogiques innovants. Thierno Tounkara a conçu le dispositif pédagogique « MindScrum » pour enseigner l'Agilité dans les projets aux élèves managers et, à ce titre, son dispositif a été labellisé par la FNEGE en 2020

Références

10 Best Data Mining Tools in 2021 (2021). URL : %5Curl%7Bhttps://monkeylearn.com/blog/data-mining-tools/%7D.

Compare DataRobot vs Dataiku DSS (2021). URL : %5Curl%7Bhttps://comparisons.financesonline.com/datarobot-vs-dataiku-dss%7D.

Data Mining Fruitful and Fun (2021). URL : %5Curl%7Bhttps://orangedatamining.com/%7D.

Plateforme professionnelle d'IA de DataRobot (2021). URL : https://www.datarobot.com/fr/.

The 16 Best Data Science and Machine Learning Platforms for 2021 (2021). URL : %5Curl%7Bhttps://solutionsreview.com/business-intelligence/the-best-data-science-and-machine-learning-platforms/%7D.

Top 15 Best Free Data Mining Tools : The Most Comprehensive List (2021). URL : %5Curl%7Bhttps://www.softwaretestinghelp.com/data-mining-tools/%7D.