M2-IA-DS

Fast Fashion & Mode Ethique

Analytic Report

Will you accepte the Javier join life classification challenge?

Pierre Matran

Résumé

Avec pour ambition construire une relation de confiance avec ses clients et de s'engager vers une approche davantage éthique du monde de la mode, la marque *Zara* a adopté une vision holistique centrée sur la protection des personnes et de la planète. L'écolabel *Join Life*, présent sur une grandes partie des articles de la marque en est l'exemple fard. Amené par un contexte ludique, l'étude présentée dans ce rapport se concentre sur l'analyse d'un jeu de donnée restreint comprenant des articles *Zara* avec leurs caractéristiques (prix, composition, présence d'écolabel). Après une étape de collecte, nettoyage et jointure des données, une analyse exploratoire a été menée afin d'acquérir une meilleure compréhension des variables disponibles et d'en extraire les informations utiles. Dans un second temps, la construction d'un modèle de classification binaire ayant pour objectif de prédire la présence ou l'absence de l'écolabel *Join Life* a été menée à bien en passant pour la l'élaboration et la comparaison de différents types de modèles naïfs, puis par l'amélioration des performances du modèle via des méthodes d'ensemble et d'optimisation d'hyperparamètres. Enfin, un outil de visualisation a été créée afin de permettre l'évaluation du modèle final grâce à un outil interactif *user-friendly*. Finalement, une interface simple et épurée a été mise à disposition dans le but de tester les prédictions modèle.

Abstract

With the ambition to build a relationship of trust with its customers and to commit to a more ethical approach to the world of fashion, the Zara brand has adopted a holistic vision focused on the protection of people and the planet. The Join Life eco-label, present on a large part of the brand's products, is a prime example. The study presented in this report focuses on the analysis of a limited data set of Zara items with their characteristics (price, composition, presence of ecolabel). After a step of data collection, cleaning and joining, an exploratory analysis was conducted in order to gain a better understanding of the available variables and to extract useful information. In a second step, the construction of a binary classification model with the objective of predicting the presence or absence of the Join Life ecolabel was carried out through the development and comparison of different types of naive models, and then through the improvement of the model's performance via ensemble methods and hyperparameter optimization. Finally, a visualization tool has been created to allow the evaluation of the final model through a user-friendly interactive tool. Finally, a simple and uncluttered interface has been made available to test the model predictions.

Table des matières

1. INTRODUCTION 4

2. CONTEXTE 6

3. DEMARCHE 6

3.1. Collecte des donnees 7
3.2. Nettoyage des donnees 7
3.3. Exploration et visualisation des donnees 7
3.4. Construction de modeles d'apprentissage 8
3.5. Communication des resultats 9

4. RESUME DES RESULTATS 9

4.1. Analyses 9
4.2. Modelisation 11

5. CONCLUSION ET PERSPECTIVES 14

Table des figures

Figure 1 - Pourcentage d'articles (non-)ecolabellises en fonction de la categorie de vetement 10
Figure 2 - Distribution et statistiques des prix des article (non-)ecolabellises. (Axe x) Prix en euros. (Axe y) Densite 10
Figure 3 - Classement des meilleurs modeles de classification binaires naïfs ranges selon l'AUC 11
Figure 4 - Classement des meilleurs modeles naïfs de classification binaire apres ajout des modeles par methodes d'ensemble 13
Figure 5 - Classement des resultats d'optimisation des hyperparametres par differents algorithmes de recherche 13

1. Introduction

L’essence de la mode réside dans l'expression de l'individualité et de la créativité. Il s’agit d’un un monde en perpétuelle évolution. Cependant, derrière cette industrie dynamique se cachent des réalités souvent troublantes.

Porteuse de valeurs intimement liées à notre épanouissement personnel, l’industrie de la mode est pourtant de plus en plus critiquée pour ses pratiques environnementales plus que problématiques. En effet, ces dernières années, les grandes entreprises de l’industrie de la mode face à la demande incessante des consommateurs pour des vêtements à la fois tendance et à des prix abordables, ont prospéré en s’appuyant sur un modèle commercial caractérise par la production rapide et à bas coût de vêtements aussi appelé *fast fashion*.

La fast fashion repose sur une chaîne d'approvisionnement mondiale complexe, où les matières premières sont extraites dans une partie du globe, transformées dans une autre, et assemblées dans une nouvelle, avant d'être finalement expédiées vers les marchés mondiaux. Cette fragmentation a permis aux grandes entreprises de la mode non seulement de réduire les coûts de production au maximum, mais également à conduire à des conditions de travail extrêmement précaires et à une exploitation accrue de la main-d'œuvre dans les pays en développement. Les travailleurs de l'industrie textile sont souvent soumis à des horaires de travail excessifs, à des salaires de subsistance et à des conditions de travail dangereuses, sans accès aux droits fondamentaux.

Au-delà du modèle économique, la *fast fashion* est responsable de la plupart des dégradations des fibres synthétiques en microparticules, d’une surutilisation de l’eau et des pesticides dans la culture du coton, d’une accumulation folle de déchets impossibles à recycler ainsi qu’à la consommation excessive d'énergie et d’émissions de gaz à effet de serre dans l’atmosphère. Ces cycles de production rapides et l'utilisation intensive de ressources naturelles épuisables sont à l’origine des conséquences particulièrement néfastes sur notre planète. De plus, la pression constante pour produire de nouvelles collections à un rythme effréné entraîne souvent des compromis en termes de durabilité et de qualité, encourageant ainsi la surconsommation et le gaspillage.

L’industrie de la mode est depuis plusieurs années sous le feu des plus vives critiques de la part des acteurs de l’action humanitaire et de l’économie sociale et solidaire. Ainsi, Oxfam, le WWF, Greenpeace et bien d’autres dénoncent, à longueur d’enquêtes et de rapports,
tous plus édifiants les uns que les autres, la tragédie sociale et environnementale qui se joue en toile de fond d’un marché dont l’accélération et la globalisation ont été véritablement spectaculaires en quelques décennies.

Pourtant, face aux incroyables défis auxquels elles sont confrontées, les grandes entreprises de la mode ont le potentiel de devenir des véritable moteurs du changement positif. En reconnaissant leurs responsabilités sociales et environnementales, elles peuvent jouer un rôle clé dans la promotion d'une mode plus éthique et durable. De nombreux leviers sont à leur disposition pour réduire leur impact, tels que l'adoption de pratiques de production plus durables, la mise en œuvre de chaînes d'approvisionnement transparentes, la promotion de

l'économie circulaire et le soutien aux initiatives locales. En investissant dans la recherche et le développement de matériaux innovants et respectueux de l'environnement, ces entreprises peuvent repenser leur approche de la mode et contribuer à la création d'un avenir plus équitable et durable.

L'écolabel *Join Life* de la marque *Zara* est l'un des exemples emblématique de ces initiatives. Ayant pour objectif de promouvoir la durabilité et de réduire l'impact environnemental de ses produits cet écolabel vise à guider les consommateurs vers des choix plus responsables en identifiant les articles de mode qui répondent à des critères spécifiques en matière de durabilité.

Les objectifs derrière la stratégie *Join Life* sont les suivants :

- ***Amélioration continue:*** *Zara* s'engage à améliorer ses pratiques de durabilité de manière continue. Cela implique d'évaluer et de revoir régulièrement les processus de conception, de production, de logistique, d'entreposage et de vente au détail pour minimiser l'impact environnemental.

- ***Modèle d'économie circulaire:*** *Zara* se tourne vers un modèle d'économie circulaire, visant à prolonger le cycle de vie des produits. Cela comprend des initiatives telles que la réparation des vêtements, la revente des articles d'occasion et les programmes de don, afin de réduire le gaspillage et de favoriser une utilisation plus durable des produits.

- **Transparence et collaboration:** *Zara* reconnaît que la réalisation de ses objectifs ne peut pas être réalisée seule. Elle travaille en étroite collaboration avec ses fournisseurs, ainsi qu'avec des organisations internationales et d'autres entreprises de l'industrie de la mode, afin de mettre en œuvre des pratiques durables à tous les niveaux de la chaîne d'approvisionnement.

- **Impact social:** Outre l'aspect environnemental, *Zara* s'engage également à améliorer les conditions sociales tout au long de sa chaîne d'approvisionnement. Cela comprend des initiatives visant à garantir des conditions de travail justes et sûres pour les travailleurs de l'industrie textile.

" We are not perfect, but we are dedicated to make things better."

ZARA

2. Contexte

Vous venez de recevoir un mail d'un ancien ami de lycée : *Javier Santos*. Vous apprenez qu'il travaille maintenant en tant que *Lead Stock Manager* dans la plus grande boutique *Zara* de Madrid. Son travaille consiste à gérer les commandes, stocks et mises rayon des articles de la marque. Il ajoute que dorénavant tous les articles possédant l'écolabel *Join Life* doivent être exposés aux clients dans un espace dédié de la boutique.

Il explique minutieusement que tous les articles leur sont livrés le lundi matin et que l'installation des étiquettes écolabellisées est à leur charge. Le bon de commande rassemble à la fois l'identifiant unique de chaque article, le type de vêtements, sa composition et s'il s'agit d'un article suivant le protocole *Join Life*.

Malheureusement, un problème informatique est survenu dans la nuit du samedi et *Javier* n'est plus en possession du bon de commande de la semaine prochaine. Il va donc lui être impossible de dissocier les articles à écolabelliser. Par chance, il possède encore le bon de commande de cette semaine sur son disque dur personnel qu'il vous a joint au mail. Dans son désespoir et désireux de conserver son tout nouveau poste de manager, *Javier* fait appel à vous pour lui venir en aide.

Fort de vos connaissances dans le domaine de l'intelligence artificielle, tout en prenant conscience de l'urgence de la situation, vous rassurez votre ami et lui proposer de jeter un œil sur les données et de tenter de construire un modèle, le plus fiable possible, pour distinguer les articles écolabellisés *Join Life* des articles classiques.

3. Démarche

Dans le cadre de ce projet d'analyse de données et de développement de modèles d'apprentissage automatique, une approche rigoureuse a été adoptée afin de garantir la transparence et la reproductibilité du travail. Dans cet esprit, l'ensemble du projet a été rendu accessible au public via un *repository* GitHub dédié. Ce dernier contient tous les fichiers et dossiers nécessaires à la compréhension et à la reproduction de la démarche. Les notebooks contenant le code source détaillé de chaque étape de l'analyse exploratoire des données (**EDA**), ainsi que les scripts pour le prétraitement des données, l'entraînement des modèles et l'évaluation des performances (**MLA**), y sont disponibles.

Toutes les manipulations et analyses de données ont été réalisées exclusivement en utilisant le langage de programmation *Python* :

- Traitements/manipulation de données : *Pandas*, *NumPy*
- Visualisation : *Seaborn*, *Matplotlib*, *Plotly*
- Apprentissage modèle : *Scikit-learn*, *PyCaret*, *LightGBM*, *CatBoost*, …
- Dashboarding/model testing : *Gradio*, *ModelExplainer*

3.1. Collecte des données

L'ensemble de données utilisés pour simuler le mail de *Javier* a été créé dans le cadre de projets de recherche par l'université de *Univeristat Oberta de Catalunya* (*Fast Fashion Eco Commitment Dataset*, *2020-11-07*) avec un objectif purement éducatif. Il est possible d'accéder à ces données via le *Hub* public appelé **Zenodo**.

DOI 10.5281/zenodo.4261101

Les données seront télécharger directement depuis le *Hub* via la bibliothèque python *zenodo_get*.

L'ensemble de données utilisés fournit des informations précieuses sur l'engagement de *Zara* en matière de durabilité environnementale. Il s'agit en réalité de 2 fichiers *csv* séparés avec les informations générales sur chaque article (identifiant, nom, description, présence d'un écolabel *Join Life*, prix, …) ainsi que leur composition (matériel, pourcentage, …).

3.2. Nettoyage des données

Afin d'obtenir des tables de données facilement utilisables, il est nécessaire de passer par une étape de nettoyage et manipulation des données brutes. Ce processus méticuleux permet d'obtenir des ensembles de données propres et prêts à être explorés, ouvrant ainsi la voie à des découvertes précises et fiables.

Les manipulations envisagées sur les données brutes sont les suivantes :

- Conversion des caractères en *utf8*, *lower case*
- Transformation des prix de centimes en euros
- Transformation des pourcentages en *float*
- Agrégation de catégories (*culotte, body, caleçon : sous-vêtement*)

Pour finir, un jointure interne (*inner join*) sera effectuée pour regrouper la table des articles ainsi que leur composition.

3.3. Exploration et visualisation des données

Une fois les données nettoyées et fusionnées correctement, il est temps de présenter de manière claire et informative les différentes caractéristiques et tendances inhérentes à l'ensemble de données. Pour cela, une serie de questions pertinentes seront traitées sur les différents aspects (distributions, relations, quantités, …) de chacune des variables :

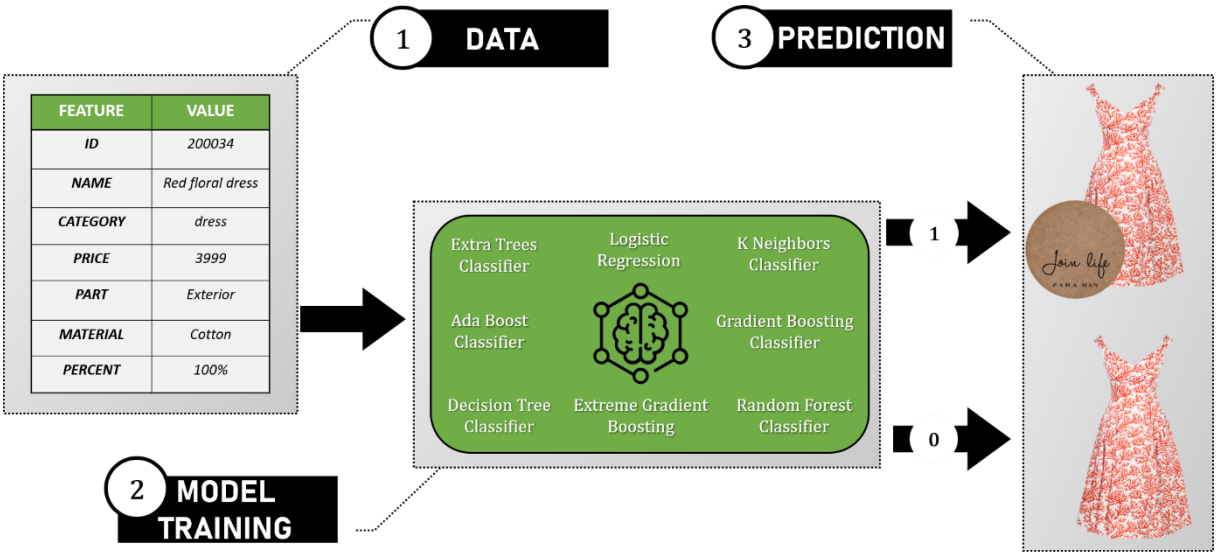
- Proportion d'écolabel *Join Life* parmi tous les articles
- Catégories de vêtements les plus susceptibles de posséder l'écolabel

- Composants des vêtements en fonction pour les doublures et le tissu principal
- Les variables les plus corrélées
- Le prix moyen en fonction des catégories de vêtement
- La distribution des prix pour les articles (non-)écolabellisés

3.4. Construction de modèles d'apprentissage

L'objectif principal est de s'appuyer sur les caractéristiques des différents articles *Zara* du jeu de donnée pour déterminer/prédire la présence (ou non) de l'écolabel *Join Life*. Il s'agit d'un processus de classification binaire où la valeur de sortie (*output*) pour un article donné est définie de la manière suivante:

- **0** : le vêtement n'est pas écolabellisé *Join_Life*
- **1** : le vêtement est écolabellisé *Join_Life*



La construction d'un modèle de classification binaire par apprentissage machine sera réaliser en suivant les étapes suivantes :

- **Feature engineering** : processus de prétraitement, sélection, extraction, normalisation et mise à l'échelle des données afin d'améliorer l'apprentissage automatique.

- **Entraînement de modèle naïfs** : processus d'entraînement et comparaison de différents modèles de *machine learning* dit « basique » c'est-à-dire avec des paramètres par défaut afin d'apprécier les modèles les plus adaptés au problème.

- **Amélioration du/des modèles** : processus au cours duquel l'objectif principale est d'améliorer les performances intrinsèques du modèle via différentes techniques (*bagging*, *boosting*, *stacking*, *blending*, optimisation des hyperparamètres, …).

- **Calibration du modèle** : processus de réajustement permettant de tendre vers un modèle probabiliste garantissant à chaque prédiction un intervalle de confiance associé.

- **Evaluation du modèle** : processus d'analyse des performances du modèles à l'aide de métriques et graphiques appropriés.

- **Finalisation du modèle** : processus de sauvegarde du modèle le plus performant en vue d'une réutilisation future en contexte de prédiction (tests, production, …).

3.5. Communication des résultats

A la suite du travail de réflexion, d'analyse, d'entrainement et de tests, il est primordial de fournir à l'utilisateur des outils simplifiés pour parcourir les résultats de l'études, la performance du modèle. Pour cela, les outils de *Dashboarding* représentent un atout majeur en offrant une visualisation claire et interactive des résultats et des métriques liés aux modèles développés. Les *dashboards* facilitent la communication avec les parties prenantes, permettent une analyse approfondie des résultats et aident à la prise de décision dite *data-driven*.

Au-delà de la performance et de l'explicabilité du modèle, il est très souvent apprécié de développer un outil basique, *user-friendly* (très généralement une interface web) qui permette de réaliser des tests de prédiction. Cela permet de mieux connaître et comprendre les flux de données gérés par le modèle (*input*, *outpout*).

4. Résumé des résultats

4.1. Analyses

L'exploration du jeu de donnée a permis de révéler de nouvelles connaissances importantes. Pour commencer, seulement *34%* des articles *Zara* disponibles possèdent un écolabel *Join life*. Certaines catégories de vêtement semblent avoir une proportion beaucoup plus importante d'article écolabellisés comme c'est le cas des tops, t-shirts et vestes (respectivement *100%, 88%, 60%*) contrairement aux sous-vêtements (à peine *7%*), ou articles de sport comme les joggings et leggings (*0%*).

Clothing categories	Not eco-labeled (%)	Eco-labeled (%)
top	0.0	100.0
camiseta	12.2	87.8
chaqueta	40.0	60.0
shorts	43.8	56.2
camisa	53.3	46.7
pantalon	60.9	39.1
falda	62.3	37.7
bermuda	70.0	30.0
vestido	74.4	25.6
blusa	80.0	20.0
underwear	93.3	6.7
jogging	100.0	0.0
legging	100.0	0.0

Figure 1 - Pourcentage d'articles (non-)écolabellisés en fonction de la catégorie de vêtement

De plus, les deux matériaux les plus utilisés pour la fabrication des vêtements *Zara* sont le polyester (*38%* partie extérieure, *77%* doublure) et le coton (*26%* partie extérieure, *13%* doublure). Á relever qu'il existe une corrélation positive entre la partie du vêtement étudié et ainsi que le taux de pourcentage des matériaux et à l'inverse une corrélation négative entre le prix des articles et la présence de l'écolabel *Join Life*. En effet, au vue de la distribution des prix des articles (non-) écolabellisés, il semble que les articles non écolabellisés se répartissent sur une gamme de prix bien plus large que les articles écolabellisés. En revanche, et contrairement aux idées reçues, leur prix moyen semble plus élevé (*~19.6€* contre *~14.5€*).

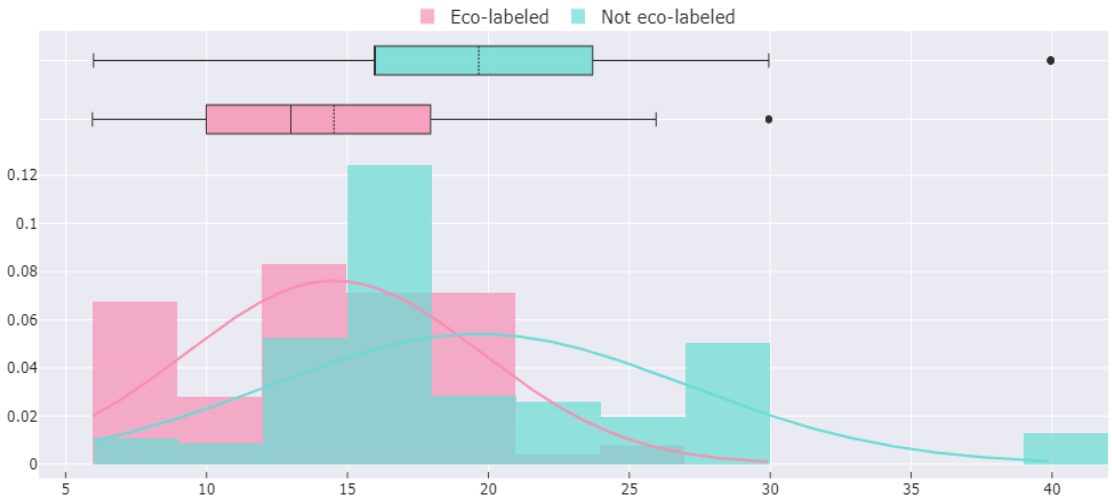


Figure 2 - Distribution et statistiques des prix des article (non-)écolabellisés. (Axe x) Prix en euros. (Axe y) Densité.

4.2. Modélisation

Á la suite du processus de *feature engineering*, de plusieurs modèles naïfs ont été construit sur une partie des initiales (*training set*). Afin de classer les modèles en fonction de leur performance, il est nécessaire de calculer certaines métriques. Dans le cas d'une classification binaire les deux métriques les plus pertinentes à évaluer sont :

- ***Accuracy*** : acuité du model, correspond au ratio des prédictions correctes sur le nombre total de prédictions faites par le modèle

- ***AUC*** *(Area Under the ROC Curve)* : capacité du modèle à faire la distinction entre les deux classes (**0** et **1**)

En réalité, l'*AUC* mesure la probabilité qu'un élément positif (**1**) choisi au hasard soit mieux classé qu'un élément négatif (**0**) choisi au hasard. Cela revient à dire que le coût d'une mauvaise classification des exemples positifs et négatifs est différent. C'est pourquoi, cette métrique semble être la mesure la plus fiable pour évaluer les performances des modèles de classification binaires de ce cas.

Les trois modèles naïfs les plus performants sont :

1. **Extreme Gradient Boosting** *(xgboost)*
2. **Random Forest Classifier** *(rf)*
3. **Light Gradient Boosting Machine** *(lightgbm)*

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.8767	0.9244	0.8462	0.8151	0.8294	0.7330	0.7343	2.3740
rf	Random Forest Classifier	0.8329	0.9164	0.7846	0.7539	0.7651	0.6359	0.6402	2.3880
lightgbm	Light Gradient Boosting Machine	0.8411	0.9139	0.8077	0.7640	0.7791	0.6558	0.6630	2.8660
et	Extra Trees Classifier	0.8521	0.9119	0.7923	0.7906	0.7883	0.6750	0.6782	2.4420
catboost	CatBoost Classifier	0.8493	0.9118	0.8000	0.7822	0.7862	0.6705	0.6752	2.8520
gbc	Gradient Boosting Classifier	0.8329	0.8962	0.7846	0.7602	0.7682	0.6381	0.6424	2.5140
ada	Ada Boost Classifier	0.7918	0.8524	0.7538	0.6942	0.7188	0.5545	0.5602	2.5880
lr	Logistic Regression	0.7370	0.8105	0.8000	0.5995	0.6839	0.4670	0.4833	5.3860
lda	Linear Discriminant Analysis	0.7123	0.8090	0.7923	0.5715	0.6623	0.4237	0.4433	2.6580
dt	Decision Tree Classifier	0.8055	0.8043	0.8000	0.6960	0.7396	0.5857	0.5953	3.9020
knn	K Neighbors Classifier	0.7014	0.7906	0.6846	0.5685	0.6200	0.3779	0.3835	2.9860
svm	SVM - Linear Kernel	0.6658	0.0000	0.6923	0.5276	0.5977	0.3220	0.3309	2.6300

Figure 3 - Classement des meilleurs modèles de classification binaires naïfs rangés selon l'AUC

Par la suite, plusieurs tentatives d'amélioration des performances des modèles vont être mises en œuvre afin de maximiser la qualité de distinction entre les articles (non-)écolabellisés :

- **Méthodes d'ensemble** : multiples « apprenant faibles » *(weak learners)* entraînés pour résoudre le même problème et combinés pour obtenir de meilleurs résultats
  - ⇨ *Bagging*
  - ⇨ *Boosting*
  - ⇨ *Stacking*
  - ⇨ *Blending*

- **Optimisation des hyperparamètres** : recherche des paramètres de modèle optimal, c'est-à-dire le jeu de paramètres pour lequel l'*AUC* est maximisée.
  - ⇨ *Random search algorithm*
  - ⇨ *Grid search algorithm*
  - ⇨ *Hyperopt algorithm*
  - ⇨ *Optuna algorithm*

A travers les différentes techniques d'amélioration du modèle, il est à noter que les méthodes de *boosting* ont davantage souffert de la quantité réduite de données en entrée par rapport aux méthodes de bagging très prometteuses à partir d'un simple modèle d'arbre de décision simple. L'assemblage et l'entraînement sur toutes les données d'entrée des 3 meilleurs modèles naïfs (*blending*) a permis d'atteindre les meilleurs résultats (modèle dit de *Voting Classifier*) avec un *AUC* à *0.9280*.

En revanche, les méthodes d'optimisation des hyperparamètres à partir de ce modèle n'ont été que très peu efficaces avec une amélioration maximale de l'*AUC* de seulement *1.6%*. Les algorithmes *Optuna* et *Hyperopt* principalement basées sur des méthodes de résolution paramétriques comme la descente de gradient ont été particulièrement décevants.

Cela peut s'expliquer en partie par l'importance d'une quantité de donnée d'entrée insuffisante et un espace de recherche extrêmement contraint par l'état initial (modèle naïf) trop proche de l'optimum, ce qui minimise les « sauts » de ces algorithmes dans l'espace des paramètres via les vecteurs propres.

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
vc	0.8575	0.9280	0.8154	0.7911	0.7999	0.6896	0.6932
lr (stacking)	0.8685	0.9273	0.8154	0.8155	0.8121	0.7113	0.7147
xgboost	0.8767	0.9244	0.8462	0.8151	0.8294	0.7330	0.7343
rf	0.8329	0.9164	0.7846	0.7539	0.7651	0.6359	0.6402
dt (bagging)	0.8438	0.9160	0.8077	0.7721	0.7855	0.6633	0.6678
lightgbm	0.8411	0.9139	0.8077	0.7640	0.7791	0.6558	0.6630
rf (stacking)	0.8740	0.9132	0.8154	0.8303	0.8188	0.7227	0.7263
et	0.8521	0.9119	0.7923	0.7906	0.7883	0.6750	0.6782
catboost	0.8493	0.9118	0.8000	0.7822	0.7862	0.6705	0.6752
lightgbm (stacking)	0.8630	0.8965	0.8000	0.8185	0.8047	0.6997	0.7039
gbc	0.8329	0.8962	0.7846	0.7602	0.7682	0.6381	0.6424
xgboost (stacking)	0.8630	0.8848	0.7692	0.8370	0.7975	0.6948	0.6998
ada	0.7918	0.8524	0.7538	0.6942	0.7188	0.5545	0.5602
lr	0.7370	0.8105	0.8000	0.5995	0.6839	0.4670	0.4833
lda	0.7123	0.8090	0.7923	0.5715	0.6623	0.4237	0.4433
dt (boosting)	0.8082	0.8081	0.8077	0.6969	0.7441	0.5921	0.6017
dt	0.8055	0.8043	0.8000	0.6960	0.7396	0.5857	0.5953
knn	0.7014	0.7906	0.6846	0.5685	0.6200	0.3779	0.3835
svm	0.6658	0.0000	0.6923	0.5276	0.5977	0.3220	0.3309

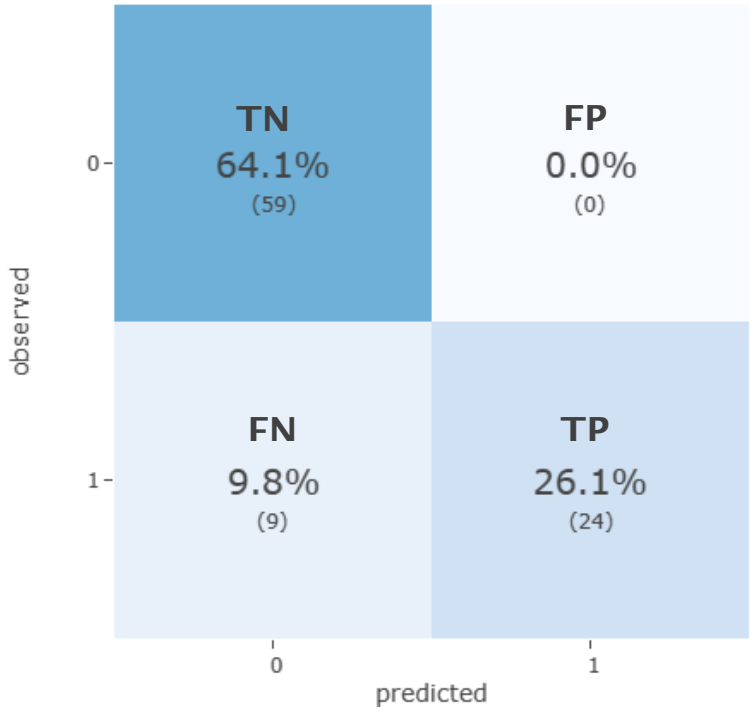
Figure 4 - Classement des meilleurs modèles naïfs de classification binaire après ajout des modèles par méthodes d'ensemble

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	CU	Accuracy_diff	AUC_diff
random	0.8685	0.9296	0.8231	0.8096	0.8141	0.7126	0.7149	1.7981	0.0110	0.0016
optuna	0.8658	0.9293	0.8154	0.8076	0.8091	0.7058	0.7083	1.7951	0.0083	0.0013
grid	0.8603	0.9288	0.8154	0.7963	0.8029	0.6950	0.6980	1.7891	0.0028	0.0008
hyperopt	0.8575	0.9295	0.8154	0.7911	0.7999	0.6896	0.6932	1.7870		0.0015
vc	0.8575	0.9280	0.8154	0.7911	0.7999	0.6896	0.6932	1.7855		

Figure 5 - Classement des résultats d'optimisation des hyperparamètres par différents algorithmes de recherche

La maximisation de l'*AUC* n'est pas la seule métrique à examiner, il est également nécessaire de contrôler et d'évaluer le comportement de modèle via plusieurs indicateurs numériques et/ou graphiques. C'est le cas par exemple de la matrice de confusion. Elle indique le nombre de vrais négatifs, ***TN*** (prédit négatif, observé négatif), de vrais positifs, ***TP*** (prédit positif, observé positif), de faux négatifs, ***FN*** (prédit négatif, mais observé positif) et de faux positifs ***FP*** (prédit positif, mais observé négatif). La quantité de faux négatifs et de faux positifs

détermine les coûts de déploiement d'un modèle imparfait en fonction d'un seuil d'acceptabilité. Le seuil optimal a été calculé à *0.74*. C'est effective au niveau de cette valeur de seuil que les nombre de ***FP*** et ***FN*** sont les plus faibles.



De nombreuses autres évaluations et graphiques sont disponibles dans les *dashboard* créé à cet effet.

5. Conclusion et perspectives

Malgré les limitations inhérentes à la quantité de données ainsi qu'aux nombres limités de variables disponibles cette étude a permis l'élaboration d'un modèle d'apprentissage automatique de classification binaire entre les articles *Zara* classiques et ceux possédant l'écolabel *Join Life*. Grâce à l'interface graphique développée, *Javier* va pouvoir dissocier les articles écolabellisés avec un précision aux alentours de *90%*.

Par ailleurs, ce modèle reste un prototype et une étude sur une base de donnée plus fournie pourrait améliorer considérablement les prédictions. Javier aurait donc tout intérêt à contacter l'équipe IT de Zara afin de leur communiquer sur l'éventualité de développement d'un outil de prédiction du même style afin d'appuyer les gestionnaires de stocks dans leur activité.