Modèles prédictifs pour la réduction des coûts associés aux non-conformités lors de la teinte des tissus

Mémoire

Catherine Bourdeau-Laferrière

Maîtrise en informatique - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

Modèles prédictifs pour la réduction des coûts associés aux non-conformités lors de la teinte des tissus

Mémoire

Catherine Bourdeau-Laferrière

Sous la direction de:

Jonathan Gaudreault, directeur de recherche
Carl Duchesne, codirecteur de recherche

Résumé

L'industrie manufacturière produit de plus en plus de produits différents avec un plus faible volume, ce qui fait considérablement augmenter la variabilité du procédé. Cette variabilité rend le maintien de la qualité des produits plus ardu qu'auparavant. L'industrie du textile ne fait pas exception à la règle. Par exemple, dans ce mémoire, le partenaire industriel, *Duvaltex* a tenté de contrôler leur non-conformité en ce qui concerne la couleur des pièces de tissu teint en mettant en place une règle d'affaire. En utilisant des données historiques fournies par l'entreprise, des modèles prédictifs ont été développés afin de pouvoir aiguiller la stratégie de test. La performance de ces modèles a été comparée à celle de leur règle d'affaires. Le modèle des forêts aléatoires améliore de 12% (taux de faux négatifs) la performance de leur règle qui était en place. Également, les modèles proposés permettent à *Duvaltex* de choisir le nombre de tests qu'ils souhaitent effectuer en fonction de leur budget ainsi que du nombre de pièces non conformes qu'ils sont prêts à tolérer.

Table des matières

Résumé ii

Table des matières iii

Liste des tableaux v

Liste des figures vi

Remerciements ix

Avant-propos x

Introduction 1

1 Notions préliminaires 3
- 1.1 Procédé de fabrication textile . . . 3
- 1.2 Évaluation de la couleur . . . 6
- 1.3 Principaux enjeux du procédé de fabrication textile . . . 7
- 1.4 Apprentissage automatique . . . 9
- 1.5 Généralisation . . . 14
- 1.6 Évaluation de la performance . . . 15

2 Problématique industrielle 17
- 2.1 Contexte industriel . . . 17
- 2.2 Objectifs . . . 19
- 2.3 Méthodologie . . . 19

3 Collecte et transformation des données 21

4 Expérimentations 31
- 4.1 180 jours vs X-jours . . . 31
- 4.2 Modèles prédictifs . . . 34
- 4.3 Impact de la méthode du partitionnement du jeu de données . . . 38
- 4.4 Impact de la taille du jeu de données d'entraînement . . . 42
- 4.5 Importance des variables . . . 46

5 Discussion 48

Conclusion **51**

Bibliographie **52**

A Article de conférence (INCOM 2021 : 17th IFAC Symposium on Information Control Problems in Manufacturing) **55**

Liste des tableaux

1.1 Type de prédiction lors de la classification binaire 16

3.1 Description des différentes tables des données sélectionnées dans *VMS* et *Laboratoire* 25
3.2 Description des différentes variables disponibles suite à l'extraction et au filtrage des données qui seront utilisées dans la construction des modèles 28

4.1 Variables utilisées dans les modèles prédictifs 35
4.2 Taux de faux négatif des modèles pour le même pourcentage de pièces de tissus testées que la *règle des 180-jours* 38

Liste des figures

1.1 Description du procédé afin de créer un tissu à partir de fils 4
1.2 Un ancien métier à tisser Jacquard. Photo libre de droit prise par George H. Williams 5
1.3 Espace de couleur CIELAB 7
1.4 Arbre de décision (Hastie et collab., 2001) 11

3.1 Diagramme d'activité UML représentant chacune des activités où des données sont générés avant, pendant ou après le procédé de teinture . . . 23
3.2 Diagramme représentant les champs qui relient chacune des différentes tables des données sélectionnées dans *VMS* et *Laboratoire* 24
3.3 Modification de la variable réponse pour tenir compte de l'ajustement de la quantité de colorant comme un signe qu'un problème a été détecté . . 29

4.1 Exactitude (Acc), taux de faux négatif (FN), taux de faux positif (FP) en fonction du nombre de jour entre deux productions du même produit pour la règle des X-jours avec un intervalle de confiance de 95%. 32
4.2 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour la *règle des X-jours* avec un intervalle de confiance de 95%. 34
4.3 Taux de faux négatif (FN) en fonction du pourcentage de pièces testées avant la production pour les trois modèles prédictifs ainsi que la *règle des X-jours* avec intervalle de confiance à 95%. 37
4.4 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *PSL* (Entraînement/Test : 75%/25%) 39
4.5 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *RL* (Entraînement/Test : 75%/25%) 40
4.6 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *FA* (Entraînement/Test : 75%/25%) 41
4.7 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentages de division du jeu d'entraînement/test avec le modèle *PLS* 43

4.8 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentage de division du jeu d'entraînement/test avec le modèle *LR* 44
4.9 Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentages de division du jeu d'entraînement/test avec le modèle *FA* 45
4.10 Importance des variables pour le modèle *FA* 46

Extraordinary claims require
extraordinary evidence

Carl Sagan

Remerciements

Premièrement, j'aimerais remercier mes directeurs de recherche, Jonathan Gaudreault et Carl Duchesne, pour leurs précieux conseils qu'ils m'ont offerts tout au long de ma maîtrise dans ce contexte particulier. Je leur suis très reconnaissante de m'avoir permis de prendre en main ce projet et d'avoir cru en mes capacités à le mener à terme malgré mon parcours plutôt atypique.

Je tiens à remercier également toutes les personnes impliquées de près ou de loin dans ce projet au sein de Duvaltex, particulièrement Sandee Noonan et Marie-Claude Côté. Sans leur temps et savoir, ce projet de recherche n'aurait pas pu être réalisé.

Avant-propos

L'article inséré en annexe a été soumis le 28 novembre 2020 et a été accepté le 6 mars 2021. L'article sera présenté dans le cadre de la conférence INCOM 2021 le 7 juin 2021. Je suis l'auteure principale de cet article.

Introduction

Dans le contexte actuel de mondialisation, l'industrie manufacturière doit constamment se renouveler afin de répondre aux demandes de haut standard de qualité de la part des clients ainsi qu'à l'augmentation de la compétition. Afin de rester compétitives, ces compagnies doivent innover pour être en mesure de répondre à la demande tout en restant efficaces (Wuest et collab., 2014).

L'industrie 4.0 a pour objectifs d'amener l'industrie à atteindre un niveau d'efficacité, de productivité et d'automatisation encore plus haut que dans les dernières révolutions industrielles (Lu, 2017). La troisième révolution industrielle a permis l'automatisation des procédés industrielles. La quatrième révoluion industrielle vise à stocker dans des bases de données et valoriser les informations concernant les caractéristiques de la matière première, le procédé ainsi que les caractéristiques des produits. Dans cette dernière révolution, on voit l'essor de techniques d'exploration des données dans ce domaine.

Malgré les défis que représente l'adaptation de ces méthodes d'exploration de données à ce domaine, leurs utilisations se multiplient grâce aux bénéfices réels qu'elles peuvent apporter. C'est le cas dans l'industrie du textile, où l'on voit les méthodes d'analyse de données utilisées dans divers contextes d'apprentissage supervisé et non supervisé. Dans le cas de l'apprentissage supervisé, on a constaté son utilisation afin de prédire des défauts de teinture du tissu (Yildirim et collab., 2018). Dans l'industrie du textile, un des paramètres de qualité les plus importants est la conformité de la couleur du tissu (Chen et collab., 2018).

La teinte des tissus industriels est un processus exigeant, tant en ce qui concerne les spécifications du client qu'en termes d'efficacité opérationnelle. De nombreux facteurs rendent difficile la reproduction exacte d'une même couleur d'un lot à l'autre. Le ma-

nufacturier doit donc faire face à de nombreuses non-conformités qui sont, malheureusement détectées seulement à la toute fin du processus de teinture, lors du contrôle de qualité.

Le but ce projet de recherche est de proposer un modèle permettant d'adapter la stratégie du contrôle qualité au sein d'une entreprise manufacturière, *Duvaltex*, afin d'identifier les pièces de tissus susceptibles d'avoir une non-conformité de la couleur avant la teinte de celle-ci.

Afin de réaliser cet objectif, des notions préliminaires sur le procédé de fabrication et de teinte du textile ainsi que des méthodes pour mesurer la couleur dans cette industrie seront détaillées dans le chapitre 1. Également, dans ce chapitre, on retrouve une revue des principaux enjeux rencontrés dans cette industrie et les techniques utilisées afin de les résoudre. Une technique utilisée dans ce projet est l'apprentissage automatique. Une brève description des deux grands types d'apprentissages ainsi que quelques algorithmes sont décrits dans le chapitre 1. Le concept de généralisation d'un modèle ainsi que différentes métriques afin de mesurer et comparer leur performance sont abordés dans ce chapitre.

Dans le chapitre 2, la problématique rencontrée chez *Duvaltex* spécifiquement est détaillée ainsi que la méthodologie pour rencontrer les objectifs fixés dans ce chapitre. Les différentes étapes afin de récupérer, nettoyer et transformer les données nécessaires à la construction des modèles prédictifs sont décrites dans le chapitre 3. Dans le chapitre 4, les expérimentations sont détaillées ainsi que les différents modèles bâtis. Finalement, au chapitre 5, une discussion sur les résultats obtenus est proposée.

Chapitre 1

Notions préliminaires

Dans ce chapitre, les notions sur lesquelles se basent les travaux réalisés dans ce mémoire sont détaillées. Dans la section 1.1, le procédé de fabrication textile est décrit. Dans la section 1.2, la méthode d'évaluation de la couleur dans l'industrie textile est expliquée. Dans la section 1.3, les principaux enjeux du procédé de fabrication textile sont énoncés. Ensuite, une revue des différents types de modèles d'apprentissage automatique est effectuée dans la section 1.4. Le concept de la généralisation d'un modèle est évoqué dans la section 1.5. Finalement, à la section 4.6, on retrouve différentes méthodes afin d'évaluer la performance d'un modèle.

1.1 Procédé de fabrication textile

Comme on peut le voir à la figure 1.1, les premiers intrants dans le procédé du textile sont les fils (1). Ceux-ci sont composés de fibres qui sont filées ensemble pour créer le fil. En fonction de la fibre utilisée, différentes méthodes de filage peuvent être employés (Hasanbeigi et collab., 2010). Il existe deux types de fibres : naturelles et fabriqués. Les fibres naturelles sont, par exemple, la laine et le coton, tandis que des fibres synthétiques couramment employées sont le polyester et la viscose (Hasanbeigi et collab., 2010). Le procédé utilisé pour passer de la fibre au fil peut avoir un impact sur l'élasticité, le volume et sa résistance à la chaleur (Hasanbeigi et collab., 2010).

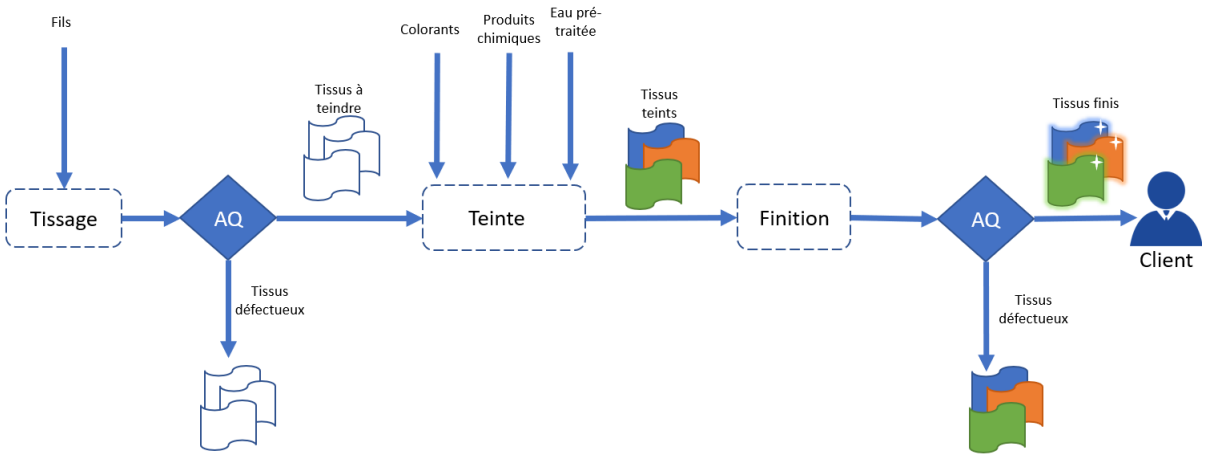

FIGURE 1.1 – Description du procédé afin de créer un tissu à partir de fils

Une fois les fils obtenus, la création du tissu peut commencer. Il existe plusieurs techniques afin de produire du tissu, mais la plus répandue est le tissage (NPTEL, 2014). Le tissage du tissu (2) s'effectue sur un métier à tisser et consiste à combiner perpendiculairement deux groupes de fils. Le groupe de fils se situant sur la longueur est appelé fils de chaîne (*warp* en anglais) et le groupe de fils orthogonal est appelé fils de trame (*weft* en anglais). Les fils sont positionnés à différent endroit en fonction du motif de tissage désiré. Le style d'un tissu est caractérisé par les fils utilisés ainsi que leur position. Un des paramètres clés dans le tissage est la tension. Si la tension appliquée sur les fils est trop grande, cela peut les briser. Par contre, si elle est trop faible, les fils peuvent se coincer dans le métier à tisser (Lee et collab., 2018).

Un des types de métiers à tisser le plus célèbres est le métier à tisser Jacquard. Inventé dans les années 1800, il y a même un lien entre l'informatique et ce métier à tisser. Malgré le fait que ce métier à tisser n'était pas alimenté électriquement, l'utilisation de cartes perforées afin de créer différents motifs sur le tissu fait en sorte que plusieurs considèrent le métier Jacquard comme le prédécesseur du programme informatique (Fernaeus et collab., 2012).

FIGURE 1.2 – Un ancien métier à tisser Jacquard. Photo libre de droit prise par George H. Williams

Comme on peut le voir sur la figure 1.1 à l'étape 3, une première inspection (l'assurance qualité (AQ)) du tissu est réalisée avant de poursuivre la production afin de s'assurer que le tissu ne présente pas de défaut. Les défauts capturés dans cette première inspection peuvent être reliés aux propriétés des fibres qui composent les fils en général, aux fils de chaîne (sur la longueur du tissu) et aux fils de trame (sur la largeur du tissu).

Le procédé de teinte de tissus, étape 4 sur la figure 1.1, est un procédé qui peut être de type discontinu, continu ou par impression (Clark, 2011). Celui utilisé par Duvaltex est le procédé discontinu. Une fois les tissus chargés dans la machine, un nettoyage de ceux-ci est effectué. Ensuite, à l'aide de colorants et de produits chimiques, le tissu est teint. Il y a deux mécanismes impliqués dans le procédé : l'absorption ainsi que la diffusion du colorant à travers les fibres du tissu. La phase d'absorption est généralement au début du procédé de teinte et se produit à des températures d'environ 30°C à 40°C (Clark, 2011). Ensuite, la température est augmentée doucement et de manière constante jusqu'à atteindre une température avoisinant les 100 °C afin de favoriser la diffusion des colorants. La température finale dépend du type de colorant utilisé. Durant tout ce processus, le tissu est aussi déplacé à une certaine vitesse dans la machine

afin d'obtenir une couleur uniforme. La durée du processus est très variable.

Ensuite, la finition du tissu, étape 5 sur la figure 1.1, varie en fonction du type de tissus final que l'on veut obtenir. Il peut être enduit d'un produit spécifique afin d'être imperméable, par exemple. Avant son inspection par l'*assurance-qualité (AQ)*, il est séché à l'aide d'un four.

C'est à l'inspection finale, étape 6 sur la figure 1.1, que la couleur du tissu est évaluée visuellement par un opérateur et/ou par spectrophotométrie. Plusieurs types de défauts peuvent être reliés à la teinte du tissu comme la non-uniformité de la couleur, la présence de taches, la déformation du tissu (pli) et la non-conformité de la couleur.

1.2 Évaluation de la couleur

L'évaluation de la couleur peut sembler être hautement subjective à la perception d'un individu. Toutefois, à travers les années, plusieurs standards se sont développés afin de pouvoir évaluer objectivement et comparer les couleurs entre elles. La convention pour mesurer la couleur utilisée par Duvaltex est le standard CIELAB développé par la Commission internationale de l'éclairage (CIE). Il est important de mentionner que ce système permet de mesurer l'apparence des couleurs et non la concentration des différents colorants impliqués dans la teinte (Weatherall et Coombs, 1992).

Comme on peut le voir à 1.3, l'espace de couleur CIELAB comporte trois dimensions : L^*, a^*, b^*. La dimension L^* décrit le degrée de luminosité d'une couleur. Les valeurs de L^* peuvent être de 0 (pour la couleur noir) à 100 (pour la couleur blanche). Les dimensions a^* et b^* représentent respectivement les couleurs opposées vert versus rouge et jaune versus bleu. Les valeurs de ces deux dimensions peuvent être positives et négatives. Plus la valeur de a^* est petite, plus la couleur a une forte composante verte. Pour des valeurs faibles sur la dimension b^*, la couleur évaluée aura une plus grande composante bleu.

L'avantage de cette convention, en plus de se rapprocher de la façon dont l'oeil humain perçoit les couleurs, est qu'elle permet de calculer la différence (ΔE) entre deux couleurs. Dans le domaine du textile, pour un même produit la différence maximale

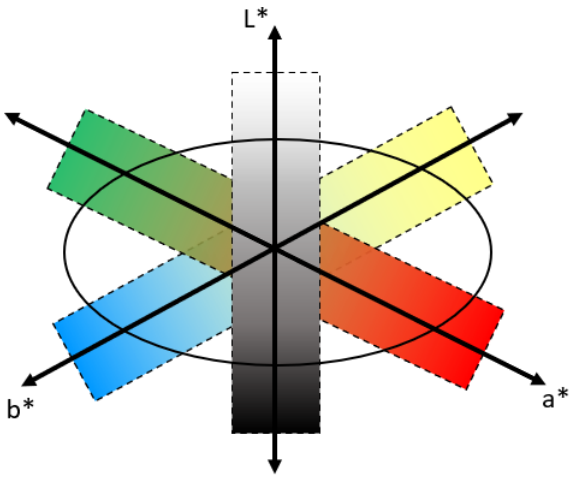

FIGURE 1.3 – Espace de couleur CIELAB

acceptée est $\Delta E = 1$, où ΔE est calculé à l'aide de l'équation 1.1.

$$\Delta E = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2} \tag{1.1}$$

Il est important que les couleurs soient mesurées avec la même source lumineuse afin de pouvoir obtenir une comparaison fiable. C'est pourquoi CIE a développé des éclairages standards afin de pouvoir utiliser efficacement leur convention d'espace de couleur. En pratique, les couleurs peuvent être évaluées à l'aide d'un spectrophotomètre spécialisé qui mesure la réflectance pour ensuite convertir les valeurs obtenues dans l'espace de couleur CIELAB (Weatherall et Coombs, 1992).

1.3 Principaux enjeux du procédé de fabrication textile

Comme dans plusieurs industries manufacturières, l'industrie du textiles rencontre des défis importants pour automatiser ou accélérer certaines opérations traditionnellement effectuées par l'humain comme l'inspection lors du contrôle qualité. Puisque les standards qualité sont très stricts, un des enjeux cruciaux est de réussir à produire des textiles dont les propriétés désirées sont atteintes à tout coup. Au niveau de la teinte du tissu, l'élaboration de la recette de coloration est un long processus manuel qui fonctionne souvent par essai-erreur et représente un défi pour les entreprises.

Plusieurs défis et enjeux du procédé de textile ont été résolus avec des approches d'apprentissage automatique et d'analyse des données. Au niveau du procédé de tissage,

l'assurance qualité est traditionnellement assurée par un humain. Cela rend le processus très subjectif, coûteux et peut entraîner des délais de production (Yildirim et collab., 2018). Afin de proposer de nouvelles approches plus automatisées de contrôle qualité, la détection automatique de défauts sur les tissus tissés a été développée en utilisant des méthodes d'apprentissage supervisées et non supervisées.

Les réseaux de neurones combinés à l'analyse en composante principale (ACP) (Su et Lu, 2011) ont été utilisés afin de détecter sept types de défauts communément rencontrés sur des tissus de lycra à partir d'image. Ils ont comparé l'efficacité d'utiliser ou non une technique de réduction de la dimensionnalité comme l'ACP avant d'utiliser un réseau de neurones pour classer les défauts. Il est intéressant de noter que la combinaison de l'ACP et du réseau de neurones donnait de meilleurs résultats en matière d'exactitude que l'utilisation seule du réseau de neurones (Eldessouki et collab., 2014). Une autre approche utilisée est l'emploi d'une caméra thermique afin de détecter les défauts en se basant sur la différence de température entre les zones défectueuses et les autres (Yıldız et collab., 2016). L'algorithme utilisé afin de procéder à la classification des défauts est la méthode du plus proche voisin (*k-nearest neighbors*).

Certains textiles plus spécialisés doivent avoir des caractéristiques très particulières afin de pouvoir être utilisés par le client. Ces caractéristiques peuvent être le pourcentage d'étirement, la charge minimale avant de déchirer et la porosité à l'air (Behera et Karthikeyan, 2006). Cependant, ces caractéristiques dépendent de plusieurs paramètres de tissage et il peut parfois être difficile d'établir les valeurs optimales de ceux-ci. Traditionnellement, ces valeurs sont obtenues à l'aide d'heuristiques ou par essai-erreur, ce qui peut rendre le développement de nouveau produit long et coûteux. Afin d'obtenir ces valeurs, des réseaux de neurones ont été utilisés et ont fourni de bons résultats.

En ce qui concerne le procédé de teinte d'un tissu, déterminer la recette de colorant à utiliser afin d'obtenir une couleur spécifique peut s'avérer une opération complexe. Certains logiciels utilisent la théorie de Kubelka-Munk afin de fournir la quantité de chacun des colorants pour obtenir une couleur désirée (Bishop et collab., 1991). Cette théorie affirme que les coefficients d'absorption et de diffusion des rayons d'une couleur résultant de l'application de plusieurs colorants sont la somme de leur coefficient d'absorption et de diffusion individuel. Cependant, cette méthode ne fonctionne pas à tout

coup. Dans la littérature, les réseaux de neurones ont été utilisés afin de formuler des recettes de coloration (Bishop et collab., 1991). Ce problème a aussi tenté d'être résolu avec des algorithmes génétiques d'optimisation (Kandi, 2007) ainsi qu'avec l'algorithme de colonies de fourmis (Chaouch et collab., 2019).

Une autre problématique du procédé de teinte est l'obtention de la couleur exacte désirée ainsi que son uniformité sur toute la grandeur du tissu. La couleur obtenue sur une pièce de tissus peut varier en fonction des propriétés de la matière première, du procédé ainsi que des propriétés du textile lui-même. Dans la littérature, des réseaux de neurones ainsi que des machines à vecteurs de support ont été utilisés afin de prédire la couleur obtenue après la teinte du tissu (Chen et collab., 2018). L'algorithme de machine à support de vecteur combiné à une optimisation par algorithme génétique a été utilisé pour prédire la conformité de la couleur ainsi que l'uniformité de celle-ci sur le tissu (Zhang et Yang, 2014).

1.4 Apprentissage automatique

Grâce à certaines percées dans les domaines des mathématiques et de l'informatique, l'intelligence artificielle gagne en popularité dans l'industrie manufacturière (Wuest et collab., 2016). La combinaison d'une plus grande accessibilité à des logiciels permettant de traiter et d'analyser des données et la quantité de données disponibles contribuent à l'essor de l'apprentissage automatique dans ce milieu (Wuest et collab., 2016). Le terme *apprentissage automatique* a été introduit pour la première fois en 1959 par A.L. Samuel dans le contexte de la réalisation d'un programme pouvant jouer aux dames et battre un humain à ce jeu (Samuel, 1959). Il a défini l'apprentissage automatique comme le principe qu'un ordinateur peut apprendre par expérience pour résoudre un problème sans avoir à le programmer spécifiquement pour la résolution de celui-ci (Samuel, 1959). Aujourd'hui, on peut diviser l'apprentissage automatique en deux grandes catégories d'algorithme : l'apprentissage supervisé et non supervisé. Cette section se concentre sur ces deux catégories de méthodes d'apprentissage. Quelques modèles appartenant à ces catégories sont présentés plus en détails.

1.4.1 Apprentissage supervisé

L'apprentissage supervisé peut se définir comme étant l'utilisation de variables X pour faire une ou des prédictions ($\hat{Y}$) d'une ou des variables Y (Hastie et collab., 2001). Les variables X et Y peuvent être quantitatives (pression mesurée en Pascal, température

mesurée en degré Celsius, longueur mesurée en mètre etc.) ou qualitative (température (chaud ou froid), longueur (petit, moyen, grand etc.). Selon le type de variables Y, différentes expressions sont utilisées pour décrire le type de modèle qui génère la ou les prédictions $\hat{Y}$. Si l'on essaie de prédire une ou des variables quantitative, on dit alors que le modèle est une *régression*. Si l'on essaie de prédire une ou des variables qualitative, on parle alors de *classification*. Certaines méthodes peuvent être plus appropriées que d'autres selon le type de variable X (Hastie et collab., 2001). L'apprentissage d'un modèle se fait à l'aide de données X et cherche à minimiser la différence entre la valeur $\hat{Y}$ et Y.

Dans cette catégorie de modèle, il y a les méthodes à base d'arbres comme les arbres de décisions. Ces méthodes font parties des cinq techniques les plus citées dans les articles concernant l'application de l'apprentissage automatique dans la planification de la production ainsi que le contrôle de procédés (Cadavid et collab., 2019). Les arbres de décisions divisent l'espace des variables X en rectangle (Hastie et collab., 2001). Prenons un exemple simple pour illustrer ce concept. Supposons que nous tentons de prédire s'il neigera. Nous avons en entrée deux variables X_1 et X_2 qui représentent respectivement la température en degrée Celsius et s'il y a des nuages dans le ciel. Au début, tout le jeu de données se situe à la *racine* de l'arbre. L'algorithme choisit une première variable pour laquelle il décide de séparer son espace de valeur par une constante qui minimise l'erreur entre $\hat{Y}$ et Y pour les observations contenues dans son jeu de données à ce *noeud*. Ici, il peut sembler logique de séparer la variable température au point 0, car les observations où il neigeait sont plus fréquentes lorsque la température est sous 0 que lorsqu'elle est au-dessus. Les observations respectant la condition illustrée sur la branche seront considérées par le prochain noeud sur cette branche. À chaque séparation, on veut s'assurer de minimiser les erreurs de prédiction. Ensuite, pour chacune des branches, l'algorithme fait un autre choix de variable et sépare à nouveau l'espace de valeur de cette variable. Le choix de la variable dans cet exemple est trivial, mais il existe plusieurs méthodes pour déterminer quelle variable choisir à chaque noeud. Si le noeud n'a pas d'autres branches, on l'appelle *noeud terminal* ou *feuille*. Ce processus continue jusqu'à ce que le critère d'arrêt soit satisfait. Le critère d'arrêt peut être le nombre d'observations minimal rencontrées dans chacune des feuilles ou la profondeur de l'arbre. Chaque feuille est représentée sous la forme de rectangle et elles représentent une prédiction comme on peut le voir à la figure 1.4.

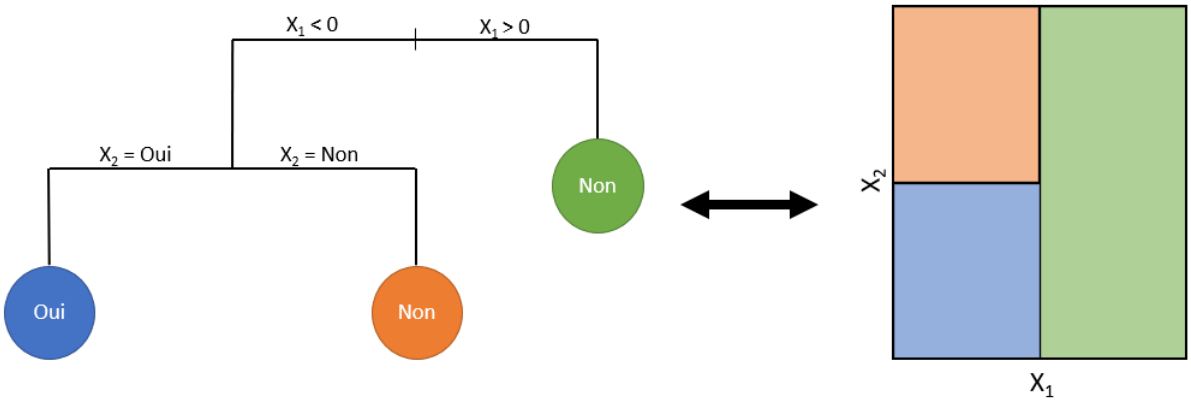

FIGURE 1.4 – Arbre de décision (Hastie et collab., 2001)

L'avantage de cette méthode est qu'elle est facilement interprétable. Elle permet aussi de traiter les valeurs manquantes sans nécessairement passer par l'imputation ni l'exclusion de l'observation en question (Hastie et collab., 2001). Cependant, les arbres de décisions sont très instables et ont une forte propension au surapprentissage (Hastie et collab., 2001). Un modèle instable ne fournit pas les mêmes prédictions alors qu'il reçoit les mêmes données en entrée. Lorsqu'un modèle est en surapprentissage, celui-ci obtient une excellente performance sur le jeu de donnée sur lequel il a été entraîné mais n'obtient pas une très bonne performance sur des données jamais vues durant l'entraînement.

Pour pallier à ce problème, les *forêts aléatoires* ont été inventés en 2001 par Leo Breiman (Breiman, 2001). Cet algorithme utilise plusieurs arbres de décision non corrélés entre eux et utilise la moyenne de leurs prédictions pour formuler une prédiction. Afin de briser la corrélation entre les arbres, les arbres de décision ne seront pas bâtis à partir des mêmes observations. En effet, une pige avec remise de n données, où n représente le nombre d'observations, est effectuée dans le jeu de données initial (Hastie et collab., 2001). Cette méthode d'échantillonnage de données est appelée *bootstrap*. Ensuite, à chaque noeud, m variables sont sélectionnées sur les p variables disponibles dans le jeu de données. Parmi ces m variables est alors choisie la meilleure variable (et valeur) pour séparer les observations. Cette étape est répétée jusqu'à l'obtention d'un noeud avec le nombre d'observations minimal. Tous ce processus sera répété b fois, où b représente le nombre d'arbres présents dans la forêt aléatoire. Les paramètres m, b ainsi que le nombre minimal d'observations sont des hyperparamètres qui sont ajustés durant l'entraînement du modèle.

La *régression logistique* fait aussi partie des méthodes les plus fréquemment utilisées dans les articles traitant de l'application d'apprentissage automatique en industrie manufacturière (Cadavid et collab., 2019). Cette méthode de classification se base sur l'estimation de la probabilité à postériori $P(C_1|n)$ qu'une observation n appartienne à une classe C_1 (Hastie et collab., 2001). Les classes C sont les différentes valeurs que peut prendre la variable Y que l'on tente de prédire. Cette probabilité est calculée à l'aide la fonction sigmoïde tel que présenté dans la formule 1.2.

$$\frac{1}{1 + exp(-a)} \quad \text{où} \quad a = w_0 + w_1 x_1 + ... + w_m x_m \tag{1.2}$$

Les coefficients w sont appelés *poids* et sont définis à l'aide des méthodes de descente du gradient et du maximum de vraisemblance (Hastie et collab., 2001). Ces poids multiplient chacune des m variables X.

Finalement, la *projection sur les structures latentes (PSL)*, une méthode combinant la régression et l'ACP, est aussi utilisée dans l'industrie manufacturière. Un des problèmes rencontrés est que les variables X peuvent être très nombreuses et corrélées entre elles. Dans un premier temps, les données utilisées sont normalisées. Pour chaque variable, la normalisation s'effectue en retirant, dans un premier temps, la moyenne des valeurs de celle-ci à chaque valeur, comme montré à l'équation 1.3. Ensuite, on divise chaque valeur par l'écart-type de toutes les valeurs centrées, comme montré à l'équation 1.4.

$$X_{k,\text{centrée}} = X_{k,\text{brute}} - \overline{X_{k,\text{brute}}} \qquad \forall\, k \;\in\; 1..K \tag{1.3}$$

$$X_{k,\text{normalisée}} = \frac{X_{k,\text{centrée}}}{\sigma_{X_{k,\text{centrée}}}} \qquad \forall\, k \;\in\; 1..K \tag{1.4}$$

Ensuite, pour ce qui est de l'ACP, on cherche à représenter le mieux possible la matrice X de dimensions $N \times K$ (où N représente le nombre d'observations et K le nombre de variables) dans un espace de dimensionnalité réduite.

$$\begin{aligned} &X = TP + E \\ &\quad \text{où } T \text{ est une matrice } N \times A \\ &\quad \text{où } P \text{ est une matrice } A \times K \\ &\quad \text{où } E \text{ est une matrice } N \times K \\ &\quad \text{où } A < K \end{aligned} \tag{1.5}$$

À l'équation 1.5, la matrice T contient les *scores* (cotes) et la matrice P contient les loadings (charges). Géométriquement, on peut interpréter la matrice P comme la matrice qui contient les vecteurs qui donnent l'orientation du nouvel espace et la matrice T contient les coordonnées des observations dans ce nouvel espace. La matrice E contient les résidus, c'est-à-dire la différence entre la matrice X et TP. Géométriquement, les résidus sont la distance de projection entre les coordonnées dans l'espace initial et le nouvel espace réduit.

En plus de réduire la dimensionnalité de la matrice X, le modèle PSL cherche à réduire la dimensionnalité de la matrice Y (variables à prédire) et créer une relation entre X et Y. La matrice Y est de dimension $N \times M$ où N représente le nombre d'observations et M le nombre de variables à prédire tel que présenté dans l'équation 1.6.

$$\begin{aligned} Y &= TC + F \\ X &= TP + E \\ T &= TW^* = TW(P'W)^{-1} \\ &\text{où } C \text{ est une matrice } A \times M \\ &\text{où } F \text{ est une matrice } N \times M \\ &\text{où } P \text{ est une matrice } A \times K \\ &\text{où } E \text{ est une matrice } N \times K \\ &\text{où } A < K \\ &W \text{ sujet à } max\ Cov(X,Y)^2 \end{aligned} \tag{1.6}$$

On obtient donc un modèle pour l'espace des Y et un modèle pour l'espace des X. La matrice T nous permet d'avoir une relation entre les deux espaces. Il est aussi important de noter que les composantes de la matrice P et W sont orthogonales entre elles. Il existe plusieurs méthodes afin d'estimer les paramètres de l'ACP ainsi que PSL comme la méthode NIPALS.

Avec des données X_{new} qui ne font pas partie du jeu de données d'entraînement, on calcule en premier la matrice T_{new} avec l'équation 1.7. Ensuite, on peut prédire les valeurs qui forme Y'_{new} avec l'équation 1.8.

$$T_{new} = X'_{new}P \tag{1.7}$$

$$Y'_{new} = T_{new}C \tag{1.8}$$

Il existe certaines mesures qui permettent de valider que les nouvelles matrices sont cohérentes avec les données du jeu d'entraînement. À l'aide de la mesure *SPE*, on peut vérifier la distance de ces nouvelles données avec les données du jeu d'entraînement. À l'aide de la mesure T^2, on peut aussi valider si la corrélation entre les variables du nouveau jeu de données est la même que celle du jeu de données d'entraînement.

1.4.2 Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne cherche pas à prédire une ou des variables Y. Le but de l'apprentissage non supervisé est de décrire les caractéristiques des observations à partir des variables X qui les caractérisent. La segmentation des données fait partie de cette catégorie. Le but de celle-ci est de trouver une façon de séparer les données en G groupes de sorte que les observations appartenant au même groupe soient plus similaires entre elles qu'avec les observations des autres groupes (Hastie et collab., 2001). Un des défis de ces méthodes est de déterminer avec combien de groupe on veut/peut séparer l'ensemble des données ainsi que la façon de mesurer la similarité des observations entre elles.

L'ACP présentée à la section précédente est un exemple d'apprentissage non-supervisé. Un autre algorithme permettant de faire de la segmentation se nomme les *K-moyennes* qui s'applique lorsque les variables X sont qualitatives et que la mesure de similarité entre les observations est une distance euclidienne (Hastie et collab., 2001). On doit déterminer préalablement combien de groupes G on souhaite obtenir. On assigne alors aléatoirement les observations à un des g groupes. On calcule alors le centre de chacun des groupes en calculant le vecteur moyen qui contient X entrées à partir des n_k observations appartenant au groupe g. Ensuite, pour chacune des observations, on calcule la distance entre chacun des centres. On assigne chacune des observations au groupe dont elle est le plus près de son centre. On répète les étapes itérativement jusqu'à ce qu'aucune observation ne change de groupe d'une itération à l'autre (Hastie et collab., 2001).

1.5 Généralisation

Le défi pour un modèle est d'obtenir une bonne capacité de *généralisation*, c'est-à-dire d'obtenir de bonnes performances de prédiction lorsqu'il est utilisé avec des données qui n'ont pas servis durant l'entraînement. On peut facilement calculer l'erreur sur le

jeu d'entraînement qui est souvent la différence absolue ou au carré entre $\hat{Y}$ et Y. Cependant, l'erreur d'entraînement est un mauvais indicateur de la capacité à généraliser d'un modèle (Hastie et collab., 2001). L'erreur d'entraînement diminue toujours lorsque la complexité du modèle augmente. Plus la complexité du modèle augmente, plus il y a un risque que le modèle surapprenne les données d'entraînement, ce qui conduira à une mauvaise généralisation du modèle (Hastie et collab., 2001). C'est pourquoi, lors de la sélection d'un modèle, pour une erreur d'entraînement égale, on privilégiera un modèle plus simple.

Afin d'estimer l'erreur de généralisation, on réserve une certaine portion, par exemple (25%) du jeu de données, appelé jeu de donnée test pour l'évaluation finale de la performance de notre modèle. On entraîne les modèles sur une certaine portion du jeu de données (50%) (jeu de données d'entraînement) et l'on compare ces différents modèles afin de sélectionner le meilleur sur une autre portion (25%) (jeu de donnée de validation) du jeu de données initial. Cependant, il n'est pas toujours facile de séparer le jeu de données en plusieurs partitions lorsqu'on n'a pas beaucoup de données. Généralement, il est plus facile de développer de bons modèles lorsqu'on a un grand jeu de données d'entraînement. Des méthodes alternatives comme la validation croisée permettent d'obtenir une bonne estimation de l'erreur de généralisation sans devoir sacrifier une partie du jeu de données afin d'estimer celle-ci.

1.6 Évaluation de la performance

L'erreur absolue ou l'erreur moyenne au carré peuvent être de bonnes métriques de performance lorsque l'on cherche à effectuer une régression. Quand est-il des métriques de performance de la classification ? Il existe différentes métriques pour mesurer la performance de modèles lors de la classification binaire. L'intérêt d'avoir plusieurs métriques afin d'évaluer la performance en classification est que parfois on veut influencer un modèle à commettre des erreurs « moins graves » que d'autres.

Pour illustrer ce propos, prenons l'exemple du développement d'un modèle visant à prédire si une personne est atteinte d'une maladie. Dans le cas où la maladie est grave, si on ne la détecte pas rapidement et que le traitement est peu envahissant, on veut favoriser un modèle qui, lorsqu'il prédit que quelqu'un n'est pas atteint de cette maladie, ce soit réellement le cas. C'est ce qu'on appelle des vrais négatifs. Puisque le traitement

est peu envahissant, peut-être que nous sommes prêts à accepter quelques prédictions qui seraient qu'une personne est atteinte, mais sans l'être réellement. C'est ce qu'on appelle des faux positifs. Dans un autre cas, si l'on cherche à développer un modèle prédisant une maladie moins grave dont le test diagnostic est très envahissant, on veut un modèle qui, lorsqu'il prédit que quelqu'un est atteint de cette maladie et doit subir le test diagnostic, ce soit vrai. C'est ce qu'on appelle de vrais positifs. Puisque le test est très envahissant et la maladie peu grave, peut-être que nous sommes prêts à accepter quelques prédictions erronées (faux négatifs) qui retournerait qu'une personne n'est pas atteinte même si elle l'est réellement. On pourra ainsi suivre l'évolution du patient et si d'autres signes de cette maladie se manifestent, décider d'effectuer le test plus invasif. Le tableau 1.1 résume les différents types de prédiction dans les cas de classification binaire.

		Y	
		0	1
$\hat{Y}$	0	Vrai Négatif	Faux Négatif
	1	Faux Positif	Vrai Positif

Tableau 1.1 – Type de prédiction lors de la classification binaire

Cependant, le jeu de données est parfois débalancé, c'est-à-dire qu'il n'y a pas toujours le même nombre d'observations dans chacune des classes. Il peut alors être plus intéressant de regarder le *taux de faux négatifs* (équation 1.9) ainsi que le *taux de faux positifs* (équation 1.10) plutôt que l'exactitude (Metz, 1978). Si notre jeu de données comportent seulement 10% d'observations négatives, notre modèle pourrait avoir une exactitude de 90% en prédisant que toutes les observations sont positives. Il ne serait d'aucune utilité pour détecter les observations négatives.

$$\text{taux de faux négatifs} = \frac{\text{faux négatifs}}{\text{nombre d'observations dont la classe est 1}} \tag{1.9}$$

$$\text{taux de faux positifs} = \frac{\text{faux positifs}}{\text{nombre d'observations dont la classe est 0}} \tag{1.10}$$

Lorsqu'on ne veut pas vraiment favoriser un type d'erreur plutôt que l'autre, il peut être intéressant de mesurer l'*exactitude* (1.11) d'un modèle.

$$\text{exactitude} = \frac{\text{vrai négatifs} + \text{vrai positifs}}{\text{nombre d'observations total}} \tag{1.11}$$

Chapitre 2

Problématique industrielle

2.1 Contexte industriel

Duvaltex est une entreprise spécialisée dans la conception et la fabrication de tissus à usage commerciale. Les tissus fabriqués sont vendus afin d'être utilisés dans la confection d'ameublements de bureau, comme recouvrement mural ainsi que dans le domaine médical. Duvaltex est un manufacturier qui possède une vaste expertise dans la chaîne de valeur de la fabrication du textile. En effet, ils possèdent la capacité de production de fil, la teinte de ceux-ci mais aussi du tissage et de la teinture en pièce ainsi que la finition de celles-ci. Duvaltex compte deux usines au Québec qui se spécialisent dans le tissage, la teinte du textile en pièce ainsi que la finition. Avec l'acquisition de True Textile en 2015, Duvaltex compte maintenant quatre usines aux États-unis également.

Chez Duvaltex, chaque production de tissage ou de teinture se voit assigner un numéro de lot. Un lot peut contenir plus d'une pièce de tissu à la fois. Aussi, à chaque lot est associé un numéro de style et de couleur. Le style est assigné au lot avant le début de la production au tissage et le numéro de couleur est associé au lot une fois que celui-ci est planifié pour être teint. La définition d'un produit est la combinaison unique d'un certain style à une couleur spécifique. Lors de la création d'un nouveau produit, une recette est créée et ajustée par les coloristes. Cette recette comprend les quantités des différents produits chimiques et colorants nécessaires à l'obtention de la couleur désirée.

Lors de la production de textile pour des clients oeuvrant dans la confection de meubles, par exemple, la couleur finale du tissu est très importante. Celle-ci doit être exactement

telle que demandée et chaque pièce d'un même lot doit avoir la même couleur finale. Aussi, la reproductibilité de la couleur est très importante, car un client qui effectue plusieurs commandes (pouvant être espacées de plusieurs semaines ou même plusieurs mois) afin de recouvrir des chaises de bureau veut être assuré que toute sa production aura la même couleur, même si les tissus utilisés proviennent de lots différents. Produire des pièces de tissus non-conformes au niveau de la couleur a un impact négatif sur la productivité. Lorsqu'il est possible de le faire, le tissu est teint une deuxième fois afin de corriger la couleur. Cela peut entraîner des retards considérables sur la livraison des commandes. Si la couleur du tissu ne peut pas être corrigée, cela cause une perte financière pour l'entreprise.

Par contre, il n'est pas si simple pour l'entreprise de s'assurer de l'obtention d'une couleur conforme à chaque pièce teinte. En effet, l'entreprise a vu son portfolio de produits augmenter drastiquement au cours des dernières années. Plutôt que de produire quelques produits en gros volume, sa production se diversifie en plusieurs produits à plus faible volume. Lors d'une étude précédente (Lajoie et collab., 2019), une corrélation entre les différentes variables du procédé (quantité d'eau utilisée, vitesse de rotation du tissus, poids du tissus, caractéristiques des colorants etc.) et la conformité de la couleur des pièces de tissus a été démontrée.

Afin d'assurer une certaine stabilité au niveau de la matière première, notamment au niveau des colorants, des efforts ont été faits afin de commander une plus grande quantité à chaque fois afin d'éviter une variabilité trop grande. Aussi, à chaque changement de lot de colorant, une analyse en laboratoire est effectuée afin de mesurer son pouvoir colorant (valeurs L*, a*, b*). Ainsi, si les caractéristiques du colorant ne sont pas les mêmes que celles du lot précédent, les coloristes peuvent ajuster les quantités utilisées dans les recettes. De plus, puisque les coloristes ont remarqué que les non-conformités de couleur semblent survenir plus souvent lorsqu'un produit n'est pas manufacturé depuis un certains temps, ils ont mis en place un protocole de tests supplémentaires. Afin de s'assurer que la recette utilisée lors du procédé de teinture donne de bons résultats, ils prélèvent un échantillon du tissu du lot à teindre et le teignent en laboratoire avec tous les produits impliqués dans le procédé à grande échelle. Cet essai est réalisé pour les produits qui n'ont pas été teints depuis plus de 180 jours. Malgré tous ces efforts, les non-conformités au niveau de la couleur se produisent toujours. Ces problématiques ne sont pas propres à Duvaltex, c'est un défi pour toute l'industrie du textile de trouver des

moyens de s'assurer de la conformité de la couleur dans un contexte ou plusieurs produits de différentes couleurs sont manufacturés avec une fréquence relativement faible.

2.2 Objectifs

Duvaltex cherche donc à améliorer sa stratégie de test afin de trouver un meilleur équilibre entre les coûts des non-conformités et le coût de la main-d'oeuvre nécessaire pour effectuer ces essais en laboratoire. L'objectif est de développer un modèle à l'aide des données historiques de production, afin de prédire si un lot est à risque d'être non-conforme au niveau de la couleur avant le début du procédé de teinte. Lorsqu'un seuil de probabilité de non-conformité est excédé, un essai en laboratoire serait réalisé sur le lot en question avant sa mise en production. La probabilité minimale déclenchant un test doit être déterminée en fonction du nombre de tests et du nombre de lots non-conformes avec lesquels l'entreprise est confortable. Ce modèle viserait à remplacer la *règle du 180 jours* qui régit présentement les essais pré-productions comme il a été expliqué dans la section 2.1. Ce modèle a pour but d'optimiser les coûts associés aux tests pré-productions en laboratoire en ciblant mieux les lots problématiques avant le début du procédé de teinte.

2.3 Méthodologie

Premièrement, une cartographie des différents systèmes d'information est effectuée afin de comprendre où se trouvent les données nécessaires au développement du modèle, sous quel format elles se trouvent et comment s'effectue leur mises à jour. À l'aide d'experts techniques chez Duvaltex, une sélection des données pertinentes est exécutée. Une fois cette étape franchie, le nettoyage et la transformation des données sont réalisés. La projection sur les structures latentes, les forêts aléatoire ainsi que la régression logistique sont utilisées afin de bâtir des modèles prédictifs sur le jeu de donnée obtenu. Ces différents algorithmes ont été choisis, car ils sont adaptés à des problèmes multivariés, comme c'est le cas dans ce contexte. Ces algorithmes proviennent aussi de différentes familles (statistiques, apprentissage automatique, etc.). Les différents modèles sont comparés entre eux ainsi qu'avec la règle du 180 jours sur la base d'indicateurs de performance. La règle du 180 jours a été transformée en règle du *X jours* afin de valider l'impact du nombre de jours sur l'efficacité à détecter des pièces de tissu pro-

blématiques. La règle du 180 jours sert donc de référence afin de comparer les modèles développés grâce aux données historiques de production.

Chapitre 3

Collecte et transformation des données

La première étape afin de débuter un projet en apprentissage automatique est d'identifier les données existantes et accessibles. Selon Tao et collab. (2018), il y a cinq types de données générées au cours d'un procédé manufacturier : données de gestion issues du progiciel de gestion intégré (PGI), données des équipements (capteurs sur les équipements), données sur les consommateurs, données du produit (produits intelligents qui auraient des capteurs intégrés) et données publiques.

Dans le cas de Duvaltex, seulement deux types de données sont disponibles : les données de gestion ainsi que les données mesurées par les capteurs placés sur les équipements. L'entreprise possède deux systèmes PGI. Les mouvements d'inventaire et les résultats des différents tests de qualité effectués sont enregistrés dans leur PGI nommé *VMS*. *Laboratoire*, leur second GPI, sert à gérer la quantité des produits (chimiques et colorants) utilisés lors de la teinte. Certains résultats de test de qualité sont capturés à l'aide d'un fichier Excel, notamment toutes les analyses du pouvoir colorant des différents lots de colorants. Pour les données captées sur les équipements, il y a les appareils de teinte et les différentes pompes des produits chimiques qui collectent de l'information. Afin de mieux comprendre comment les différents systèmes interagissent entre eux et à quelles étapes les données sont générées, un diagramme d'activité a été créé (figure 3.1).

Comme on peut le voir à la figure 3.1, la pièce de tissu apparaît pour la première fois dans le système *VMS* lorsqu'une commande est assignée au tissage. Ensuite, elle se voit assigner un numéro de style ainsi que toutes les caractéristiques associées à sa production au tissage. Ensuite, elle se voit assigner une date de réception à l'usine de

St-Victor (STV) où le procédé de teinte se déroule. Une fois sa production planifiée, un numéro de lot de teinture est assigné à la pièce ainsi que le numéro de l'équipement sur lequel sa teinte se produira. Avant de lancer la production, les paramètres de production sont révisés dans *Laboratoire*. Lors de la production, différents paramètres (température, pression, vitesse de rotation, niveau, etc.) sont enregistrés dans l'automate de la machine de teinte. Ces données sont ensuite archivées et accessibles dans un serveur sous forme d'un fichier de base de données. Un fichier par lot de production est généré. Parallèlement, les informations sur les quantités de colorant ainsi que celles des produits chimiques sont enregistrées à la fin de la production. Certaines informations sont mesurées en parallèle du procédé de teinte, comme les pouvoirs colorant de chacun des colorants reçus à STV, la qualité de l'eau ainsi que l'analyse des fils reçus.

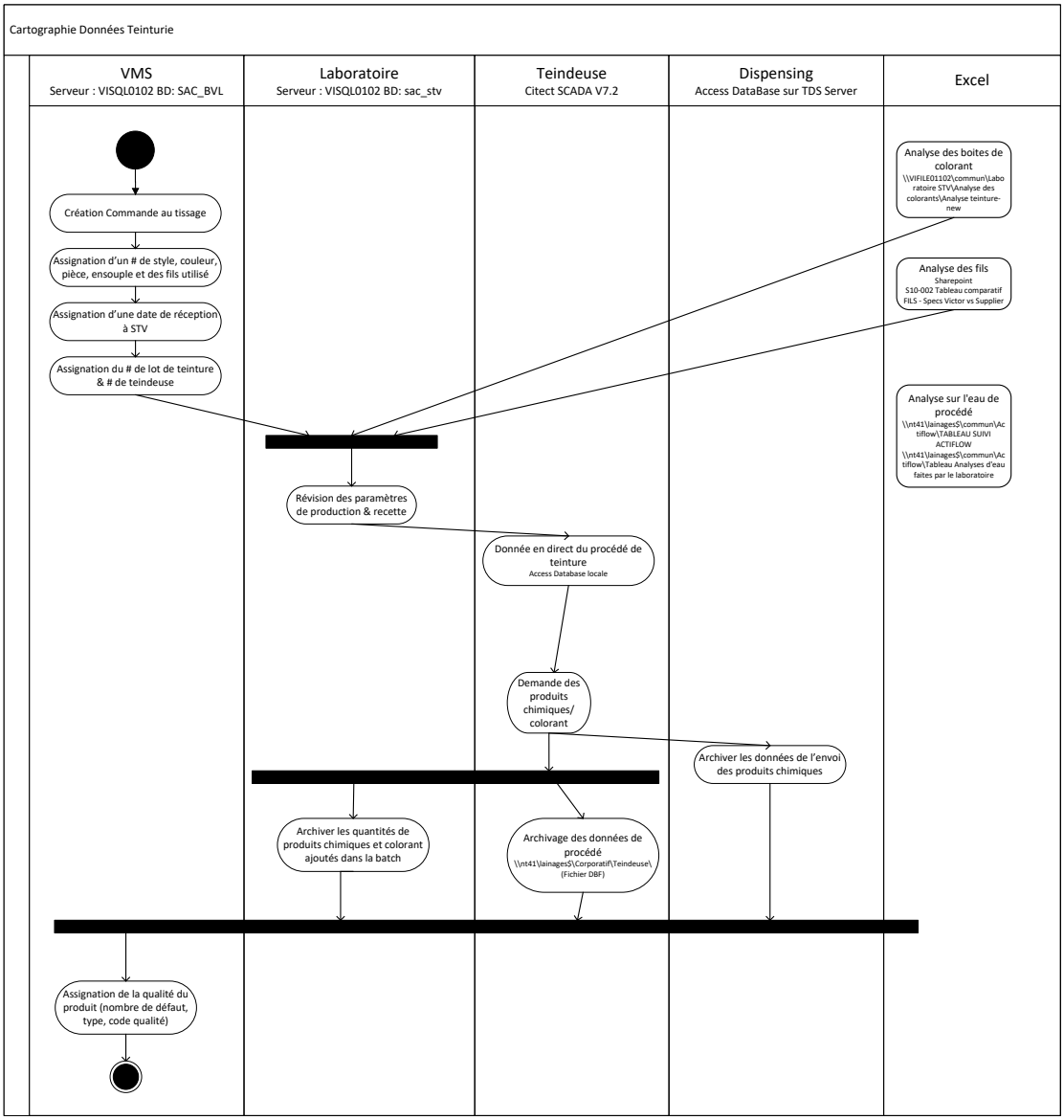

FIGURE 3.1 – Diagramme d'activité UML représentant chacune des activités où des données sont générés avant, pendant ou après le procédé de teinture

Suite à cette cartographie des processus informationnels, l'extraction des données des différentes sources a été effectuée rétroactivement à compter du début janvier 2018 jusqu'à mars 2020. Puisque l'information est extraite de plusieurs sources différentes, leur contenu a été analysé afin de comprendre comment les relier entre elles. À la figure 3.2, on retrouve les différents liens entre chacunes des tables d'information. Une fois les tables reliées entre elles, les champs d'intérêt ont été extrait. Le contenu pertinent par table est décrit au Tableau 3.1.

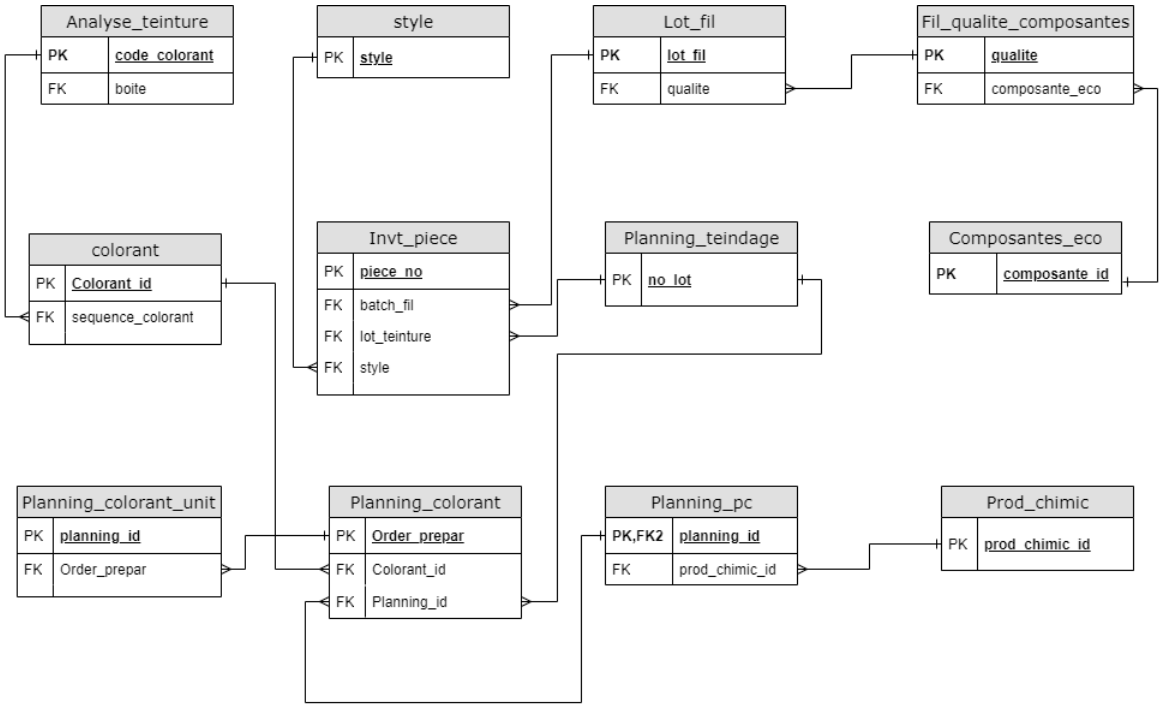

FIGURE 3.2 – Diagramme représentant les champs qui relient chacune des différentes tables des données sélectionnées dans *VMS* et *Laboratoire*

Table	Description
Analyse_teinture	Contient les informations sur les valeurs L*, a*, b* des colorants.
Style	Contient les caractéristiques (nombre de duites, nombre de brins et poids par verge tissée) du tissage des différents styles.
Lot_fil	Associe le numéro de lot du fil avec son code de type de fil.
Fil_qualite_composantes	Associe le type de fil avec le pourcentage de ses composantes
Colorant	Associe le code du colorant avec sa description.
Invt_piece	Contient tous les changements de statut du tissus avec les dates auxquelles ceux-ci se sont produits. Contient le numéro de lot de tous les fils utilisés lors du tissage ainsi que le numéro de lot associé au procédé de teinte. Contient le numéro de style ainsi que celui de la couleur associée au tissus.
Planning_teindage	Contient la date de la mise en production de la teinte ainsi que le poids de la pièce teinte.
Composantes_eco	Donne les composantes des fibres que l'on peut retrouver dans les fils (fibres vierges, recyclées etc.).
Planning_colorant_unit	Associe le numéro de lot de teinte aux numéros de lot des colorants utilisés durant cette production.
Planning_colorant	Contient la quantité de colorant requise pour chacun des lots de production.
Planning_pc	Contient la quantité utilisée de chacun des produits chimiques impliqués dans le procédé de teinte pour chaque lot de production.
Prod_chimic	Associe le code du produit chimique à sa description.

Tableau 3.1 – Description des différentes tables des données sélectionnées dans *VMS* et *Laboratoire*

Une fois les tables de données décrite dans le Tableau 3.1 intégrées ensemble, un nettoyage de celles-ci a été effectué. Seulement les données fournissant de l'information sur les tissus qui ont été teints ont été conservées. Les doublons et les entrées n'ayant pas de numéro de lot de production ont été retirés. Aussi, les données concernant des productions antérieures au 1er janvier 2019 ont été retirées. Cette décision a été prise, car beaucoup d'améliorations ont été apportées au procédé ont été apportées depuis ce temps et cela aurait entraîné beaucoup d'informations non-pertinentes dans les données. Le Tableau 3.2 contient toutes les variables pertinentes des tables de données du Tableau 3.1 ainsi que leurs descriptions. Après avoir examiné les données de ces

différentes variables, les données portant sur les produits chimiques ont été écartées de l'analyse, car la valeur de la recette pour un même produit était toujours la même et cela n'apportait pas d'information supplémentaire. Pour différents produits, il était compliqué d'utiliser ces informations, car il existe une panoplie de produits chimiques et l'information sur ceux-ci est très limitée dans la base de données (seulement le nom y était disponible). De plus, certains produits chimiques avaient changé de nom et de code dans la base de donnée, même si c'était le même produit chimique, ce qui rendait difficile l'analyse de l'impact de ceux-ci.

Variable	Type	Description
Style	Qualitative Nominale	Numéro unique correspondant au style de tissage.
Color	Qualitative Nominale	Numéro unique correspondant à la couleur lors du procédé de teinte.
Piece_no	Qualitative Nominale	Numéro unique correspondant à la pièce de tissu.
date_produced	Qualitative Ordinale	Date associée au changement de status(prod_status).
prod_status	Qualitative Nominale	Indique à quelle étape du procédé la pièce de tissu se retrouve. Ce qui nous intéresse ce sont les pièces dont le status est à VV (Vente) ou à 100 (Inventaire).
lot_teinture	Qualitative Nominale	Numéro unique associé à la mise en production du procédé de teinte.
batch_fil_1 à 10	Qualitative Nominale	Numéro unique du lot de fil utilisé à différentes positions durant le tissage.
description_en	Qualitative Nominale	Type de fibres (recyclés, vierge etc).
pourcentage	Quantitative continue	Pourcentage de chaque type de fibres dans les fils.
width	Quantitative Continue	Largeur de la pièce de tissu.

Variable	Type	Description
piece_bebe	Qualitative Booléenne	Indique si c'est la pièce de tissu originale (N) ou si c'est une pièce provenant d'une autre (Y). Afin d'éviter les doublons, c'est les informations sur les pièces originales qui seront retenues.
quality_code	Qualitative Nominale	Indique si le tissu est conforme ou non. Les tissus ayant une non-conformité suite au procédé de teinte sont associés au code 23.
quality_sous_code	Qualitative Nominale	Indique quel type de non-conformité au niveau du processus de la teinte est associé au tissu. Les problèmes de couleurs non-conformes sont associés au code 61.
debut_teindage	Qualitative Ordinale	Indique la date et heure à laquelle le procédé de teinte a débuté.
poids_total_pieces	Quantitative continue	Indique le poids de la pièce avant la teinte de celle-ci.
machine_id	Qualitative Nominale	Numéro de la machine avec laquelle la pièce de tissu a été teinte.
duites_tissees	Quantitative Continue	Nombre de duites du tissu
brins	Quantitative Continue	Nombre de brins du tissu.
once_verge_tissee	Quantitative Continue	Poids par verge de tissus
description_colorant	Qualitative Nominale	Nom du colorant utilisé
qty_requise_colorant	Quantitative Continue	Quantité de colorant dans la recette de teinte
L	Quantitative Continue	Valeur colorimétrique "L*" du colorant.
A	Quantitative Continue	Valeur colorimétrique "a*" du colorant.
B	Quantitative Continue	Valeur colorimétrique "b*" du colorant

Variable	Type	Description
description_pc	Qualitative Nominale	Nom du produit chimique utilisé
qty_utilisation_pc	Quantitative Continue	Quantité de produit chimique utilisée.

Tableau 3.2 – Description des différentes variables disponibles suite à l'extraction et au filtrage des données qui seront utilisées dans la construction des modèles

Une des hypothèses formulées par les experts chez Duvaltex était que le changement de lot de fils d'un lot de production de tissus à l'autre semblait causer des non-conformités au niveau de la couleur après le procédé de teinte. Afin de capturer ces changements de lot, des nouvelles variables qualitatives binaires ont été créées : Changement_de_fil 1 à 10. Pour la production d'un lot donné, si le numéro de lot pour le fil à la même position a changé depuis la dernière production pour ce même produit alors la nouvelle variable indiquera 1 sinon, ce sera 0. Les experts soutenaient aussi que les changements de fils n'ont pas tous le même impact sur la qualité. Les changements de fils composés de fibres recyclées seraient plus problématiques, car leur composition risque de changer beaucoup plus d'un lot à l'autre que lors d'un changement de fils composés de fibres vierges. Pour tenir compte de cette information, quatre variables par position de fil ont été créées : Changement_fil_PC, Changement_fil_PI, Changement_fil_Virgin, Changement_fil_Eco_Poly. Ci-dessous, on retrouve le détails des calculs de ces quatre variables (équation 3.1 à 3.4).

$$Changement\_fil\_Eco = \frac{\sum_{n=1}^{8} Changement\_fil\_n \times pourcentage\_Eco\_n}{nombre\_fil} \quad (3.1)$$

$$Changement\_fil\_PI = \frac{\sum_{n=1}^{8} Changement\_fil\_n \times pourcentage\_PI\_n}{nombre\_fil} \quad (3.2)$$

$$Changement\_fil\_PC = \frac{\sum_{n=1}^{8} Changement\_fil\_n \times pourcentage\_PC\_n}{nombre\_fil} \quad (3.3)$$

$$Changement\_fil\_Virgin = \frac{\sum_{n=1}^{8} Changement\_fil\_n \times pourcentage\_Virgin\_n}{nombre\_fil} \quad (3.4)$$

Pour ce qui est des colorants, la variable représentant le nom de ceux-ci n'a pas été utilisé. Puisque l'information concernant leur pouvoir colorant (L*, a*, b*) était disponible, seulement ces variables ont été retenues. Cela veut dire que pour un produit ayant une recette à 6 colorants, 18 variables caractérisent ceux-ci. Afin de pouvoir mieux capturer les changements de lot de production pour le même produit, les différences

entre les valeurs colorimétriques pour les mêmes colorants ont été quantifiées à l'aide de trois variables : diff_L*, diff_a* et diff_b* (équation 3.5 à 3.7).

$$diff\_L* = \sum_{n=1}^{6}(colorant\_n\_L*_{i+1} - colorant\_n\_L*_{i})^2 \tag{3.5}$$

$$diff\_a* = \sum_{n=1}^{6}(colorant\_n\_a*_{i+1} - colorant\_n\_a*_{i})^2 \tag{3.6}$$

$$diff\_b* = \sum_{n=1}^{6}(colorant\_n\_b*_{i+1} - colorant\_n\_b*_{i})^2 \tag{3.7}$$

On cherche ultimement à prédire la probabilité qu'un produit soit conforme ou non au niveau de la couleur. La non-conformité de la couleur correspond à tous les tissus dont les variables *quality_code* et *quality_sous_code* possèdent les valeurs 23 et 61 respectivement. Suite aux discussions avec l'équipe qualité au laboratoire, il a été soulevé que des ajustements peuvent être fait à la quantité de colorant lorsqu'on se rend compte qu'il y a un risque de problème au niveau de la teinte causé soit par un changement des pouvoirs colorants ou après avoir observé une différence de couleur de base avec le tissu après le tissage. Afin de tenir compte de ces problèmes dans le modèle, on a modifié la variable à prédire pour détecter non seulement les tissus non-conformes au niveau de la couleur, mais aussi ceux dont les quantités de colorant dans leur recette ont été modifiées significativement. La nouvelle variable *IsProblem* retourne la valeur 1 pour les cas en rouge dans la figure 3.3. Sinon, elle prend la valeur zéro.

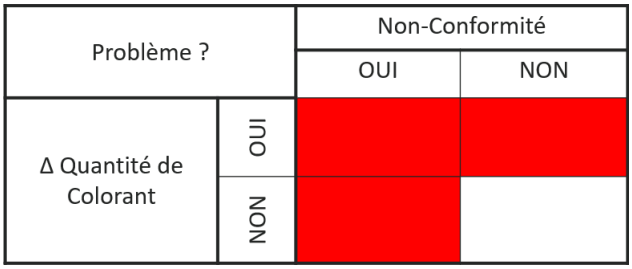

FIGURE 3.3 – Modification de la variable réponse pour tenir compte de l'ajustement de la quantité de colorant comme un signe qu'un problème a été détecté

Afin de quantifier les changements de quantité de colorant d'une production à l'autre pour le même produit, il est important de considérer le poids du tissu teint. La quantité de colorant augmente avec le poids de la pièce et n'est pas nécessairement le signe qu'un

ajustement a été fait à la recette. L'équation 3.8 présente le calcul de cette nouvelle variable.

$$\text{diff\_col\_n} = \frac{qty\_requise\_colorant_i}{poids\_total\_pieces}_{i+1} - \frac{qty\_requise\_colorant_i}{poids\_total\_pieces}_{i} \quad \forall\, n \in \{1, ..., 6\} \tag{3.8}$$

Ensuite, afin de pouvoir comparer toutes les productions ensemble, un ajustement par rapport à la quantité de colorant total utilisée a été fait pour obtenir la variable *Diff_ col* qui exprime le changement global dans la recette de colorant. L'équation 3.9 présente la formule utilisée pour calculer cette variable.

$$\text{Diff\_col} = \frac{\sum_{n=1}^{6} diff\_col\_n^2 \times qty\_requise\_colorant_i}{\sum_{n=1}^{6} qty\_requise\_colorant_i} \tag{3.9}$$

Afin de déterminer la valeur pour laquelle la variable *Diff_ col* correspond à un changement significatif dans la recette, plusieurs valeurs ont été testées dans la phase d'entraînement des modèles afin de déterminer laquelle donnait les meilleurs résultats.

Également, la variable *diff_ days* a été créée afin de mesurer le nombre de jours qui s'est écoulé depuis la dernière production du même produit (valeurs pour les variables *style* et *color* égales).

Chapitre 4

Expérimentations

La section suivante décrit les différentes expérimentations effectuées afin de mesurer, dans un premier temps, l'efficacité de la *règle des 180 jours* établie par *Duvaltex* pour prévenir les non-conformités en ce qui concerne la couleur. Ensuite, l'efficacité de trois modèles prédictifs a été mesurée : les forêts aléatoires, la régression logistique ainsi que la projection sur les structures latentes.

4.1 180 jours vs X-jours

Premièrement, l'efficacité à détecter les tissus non conformes de la *règle des 180 jours* a été évaluée en utilisant les données historiques disponibles. Afin de quantifier l'impact du nombre de jours sur la conformité de la couleur, on peut questionner si 180 jours et plus est vraiment le bon intervalle de temps pour détecter un maximum de non-conformités. Cette règle a donc été généralisée pour devenir la *règle des X-jours* où différents intervalles de temps (1 à 500 jours) ont été testés. La performance de la *règle des X-jours* a été mesurée en utilisant les mesures suivantes : taux de faux négatifs (FN), taux de faux positifs (FP) ainsi que l'exactitude (Acc).

Comme on peut le voir à la figure 4.1, pour la *règle du 180 jours*, on obtient une exactitude de 73,2%, un taux de faux négatif de 92,8% et un taux de faux positif de 6,6%. On remarque à figure 4.1 que lorsque FN augmente FP diminue. Le but est de sélectionner l'intervalle de temps qui donne le plus petit FP et FN possible. Il est aussi important de noter que pour Duvaltex, la mesure la plus importante est de réduire le nombre de FN. Cette mesure est la plus importante, car un faux positif est moins coûteux qu'un faux négatif pour eux.

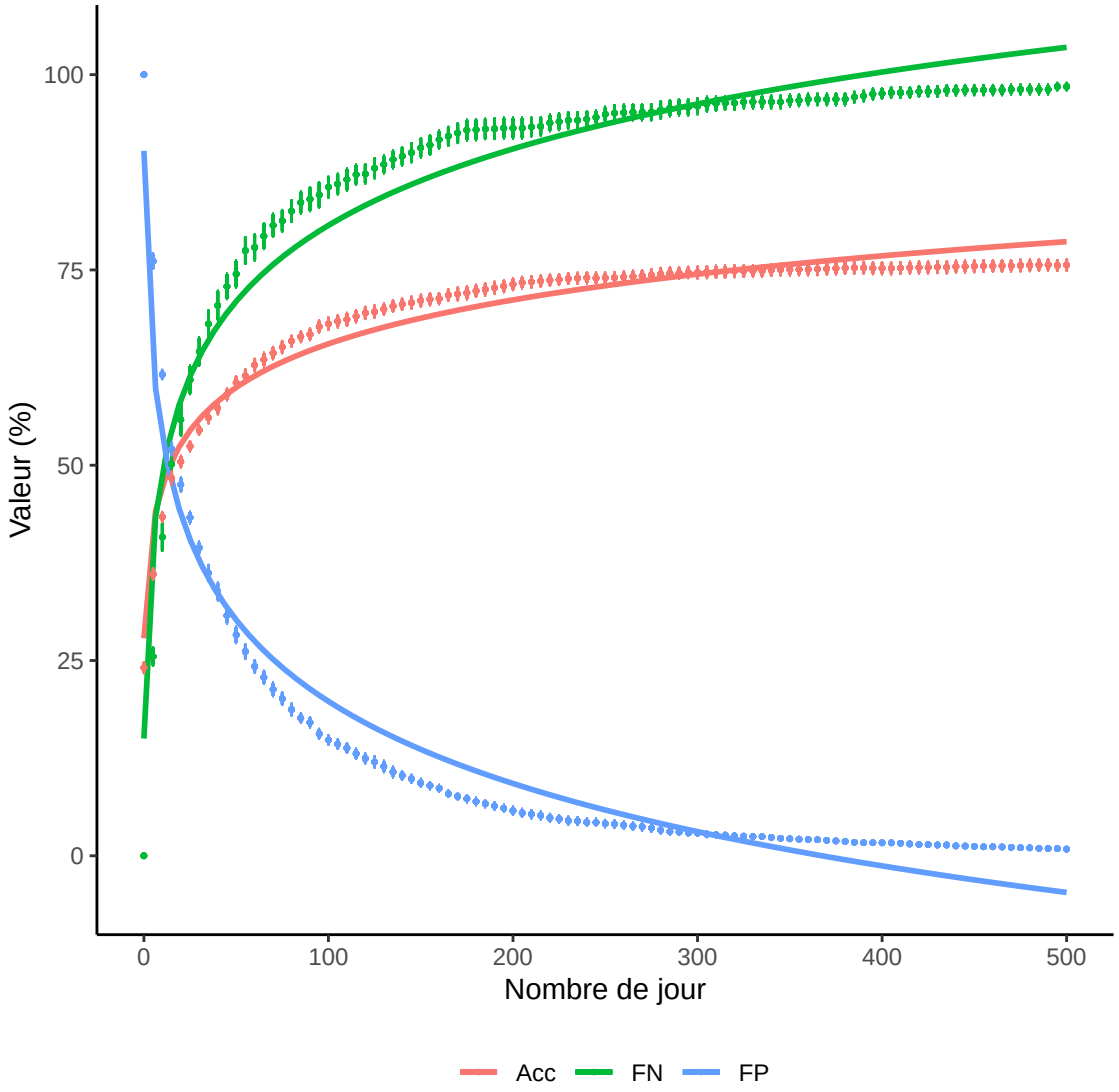

FIGURE 4.1 – Exactitude (Acc), taux de faux négatif (FN), taux de faux positif (FP) en fonction du nombre de jour entre deux productions du même produit pour la règle des X-jours avec un intervalle de confiance de 95%.

Plutôt que d'utiliser l'intervalle de temps entre deux productions du même produit sur l'abscisse, ce chiffre a été converti en pourcentage de pièces testées avant la production. Plus l'intervalle de temps est petit, plus la quantité de tests à effectuer sera grande. De cette façon, *Duvaltex* peut constater comment leur stratégie de test influence le nombre de non-conformités qui ne seraient pas détectées. On constate, à la figure 4.2, que le FN diminue lorsque le pourcentage de tissus qui doivent être testés avant d'être teints

augmente. Si on effectue des tests sur chacune des pièces de tissu avant d'être teinte, le taux de faux négatif serait de 0%. Cependant, ce n'est pas une approche envisageable pour la compagnie, car d'un point de vue de main-d'œuvre, ce ne serait pas réalisable de tester toutes les pièces de tissu avant le procédé de teinte. L'avantage de cette visualisation est de permettre à *Duvaltex* d'établir un budget de test en tenant compte du coût de ne pas détecter une pièce de tissu problématique versus le prix d'effectuer le test avant de mettre la pièce de tissu en production. La relation qu'on observe entre le nombre de tests effectués et le taux de faux négatifs est linéaire, ce qui n'est pas très performant comme modèle. En effet, il serait plus avantageux d'avoir une relation non-linéaire, logarithmique par exemple, car pour chaque test supplémentaire effectué, on réduirait de manière beaucoup plus drastique le taux de faux négatifs qu'avec une relation linéaire.

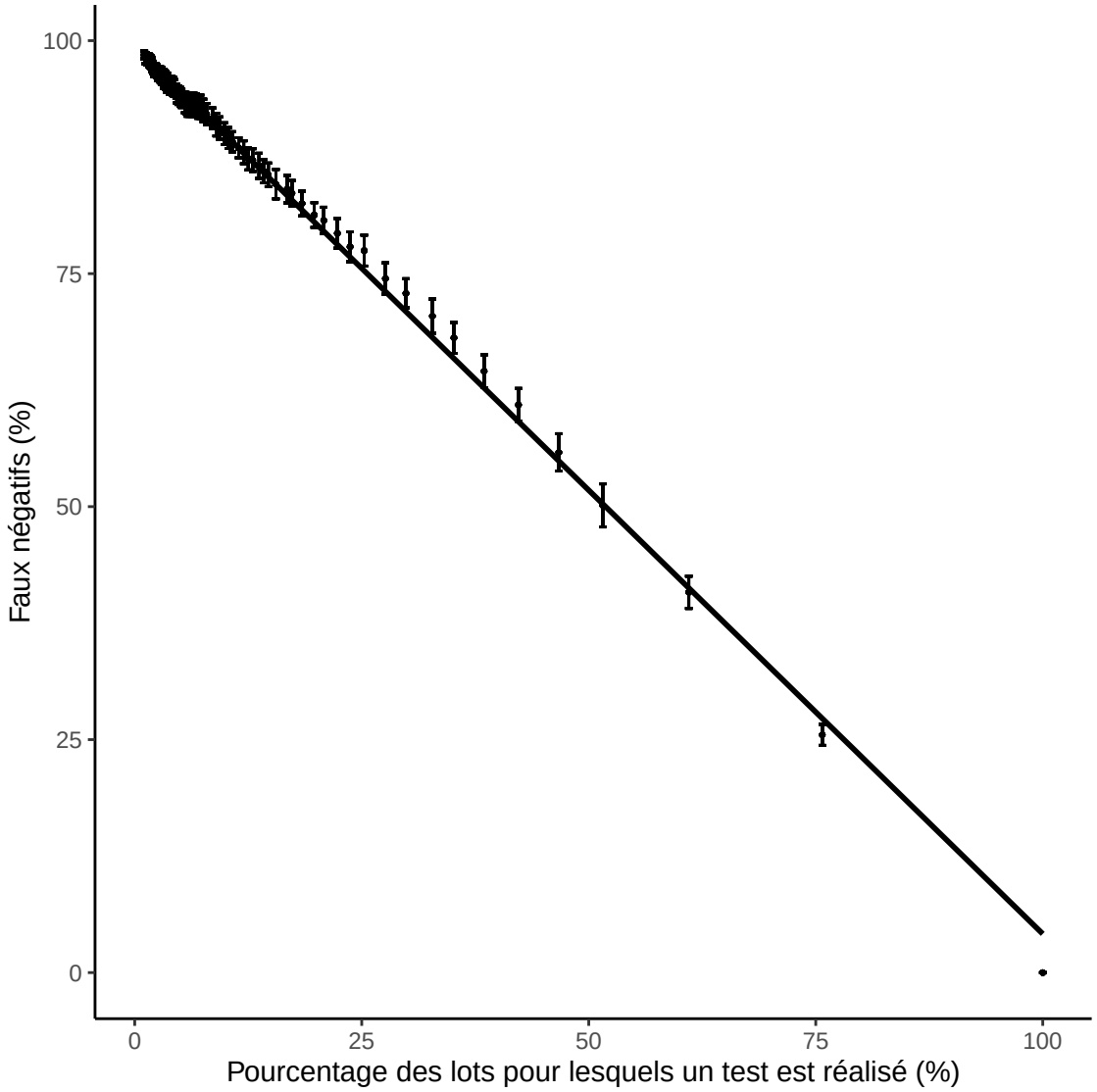

FIGURE 4.2 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour la *règle des X-jours* avec un intervalle de confiance de 95%.

4.2 Modèles prédictifs

De toutes les variables disponibles, il y a douze variables (présentées au tableau 4.1) qui ont été retenues afin de bâtir les modèles prédictifs, car les autres variables n'amélioraient pas la performance des modèles. On peut regrouper ces variables en trois catégories : caractéristiques du produit, caractéristiques du lot au tissage, caractéristiques du lot à la teinte. Pour la première catégorie (1), les valeurs de ces variables restent les mêmes à travers les différentes productions et sont propres au produit lui-même.

La deuxième catégorie (2) de variables possède des valeurs qui fluctuent à travers les productions. Finalement, la dernière catégorie de variables (3) possèdent également des valeurs qui fluctuent à travers les productions.

Catégorie	Nom de la variable
Caractéristiques du produit (1)	duites_tissées
	brins
	Nombre de fils
	Nombre de colorant
Caractéristiques du lot au tissage (2)	Changement_fil_Eco
	Changement_fil_PI
	Changement_fil_PC
	Changement_fil_Virgin
	poids_total_pieces
Caractéristiques du lot à la teinte (3)	diff_L
	diff_a
	diff_b
	diff_days

Tableau 4.1 – Variables utilisées dans les modèles prédictifs

Tout d'abord, le jeu de données a été séparé chronologiquement. Les données les plus anciennes ont été utilisées (75% du jeu de données entier) pour l'entraînement des modèles. Un échantillon de 25% du jeu de données entier a été réservé et non utilisé lors de l'entraînement des modèles afin d'évaluer la performance de ceux-ci.

Trois modèles ont été entraînés avec ce jeu de données : la forêt aléatoire (FA), la régression logistique (RL) et la projection sur les structures latentes (PSL). Les hyperparamètres suivants du modèle de forêt aléatoire ont été ajustés en utilisant le jeu de données d'entraînement : le nombre d'observations minimum dans les noeuds terminaux et le nombre de variables à utiliser pour effectuer une séparation à chaque noeud. Pour le nombre d'observations minimum à conserver dans les noeuds terminaux, des valeurs entre 2 et 10 ont été testées et c'est la valeur de 4 qui a été sélectionné. Pour le nombre de variables à utiliser à chaque noeud, des valeurs entre 1 et 4 ont été testées et c'est la valeur de 3 qui a été retenue. Le nombre d'arbres dans la forêt aléatoire a été fixé à 500 arbres. Pour le modèle *PSL*, le nombre de composantes a été fixé à 10, car c'est le nombre de composantes avec lequel les meilleurs résultats de classification ont été

obtenus en entraînement.

La performance de ces modèles est comparée à celle de la *règle des X-jours*. Les modèles retournent pour chacune des observations du jeu de données test une valeur continue entre 0.0 et 1.0. Cette valeur peut être interprétée comme étant la probabilité de la pièce de tissu d'être non conforme en ce qui concerne la couleur. On peut sélectionner un seuil (entre 0.0 et 1.0) pour lequel on considère qu'on doit tester la pièce avant de débuter le procédé de teinte sur celle-ci. Pour chacune des valeurs possibles de ce seuil, on peut en tirer le pourcentage des lots de production chez Duvaltex qui devraient être testés. C'est donc cette dernière valeur qu'on utilisera sur l'axe des abscisse de la figure 4.3. Plus petit ce seuil est, plus grand est le nombre de lots à tester par l'équipe de laboratoire. Sur la figure 4.3, le pourcentage de tests en préproduction sur l'abscisse est obtenu en faisant varier le seuil de prédiction pour lequel on considère une pièce de tissu non conforme.

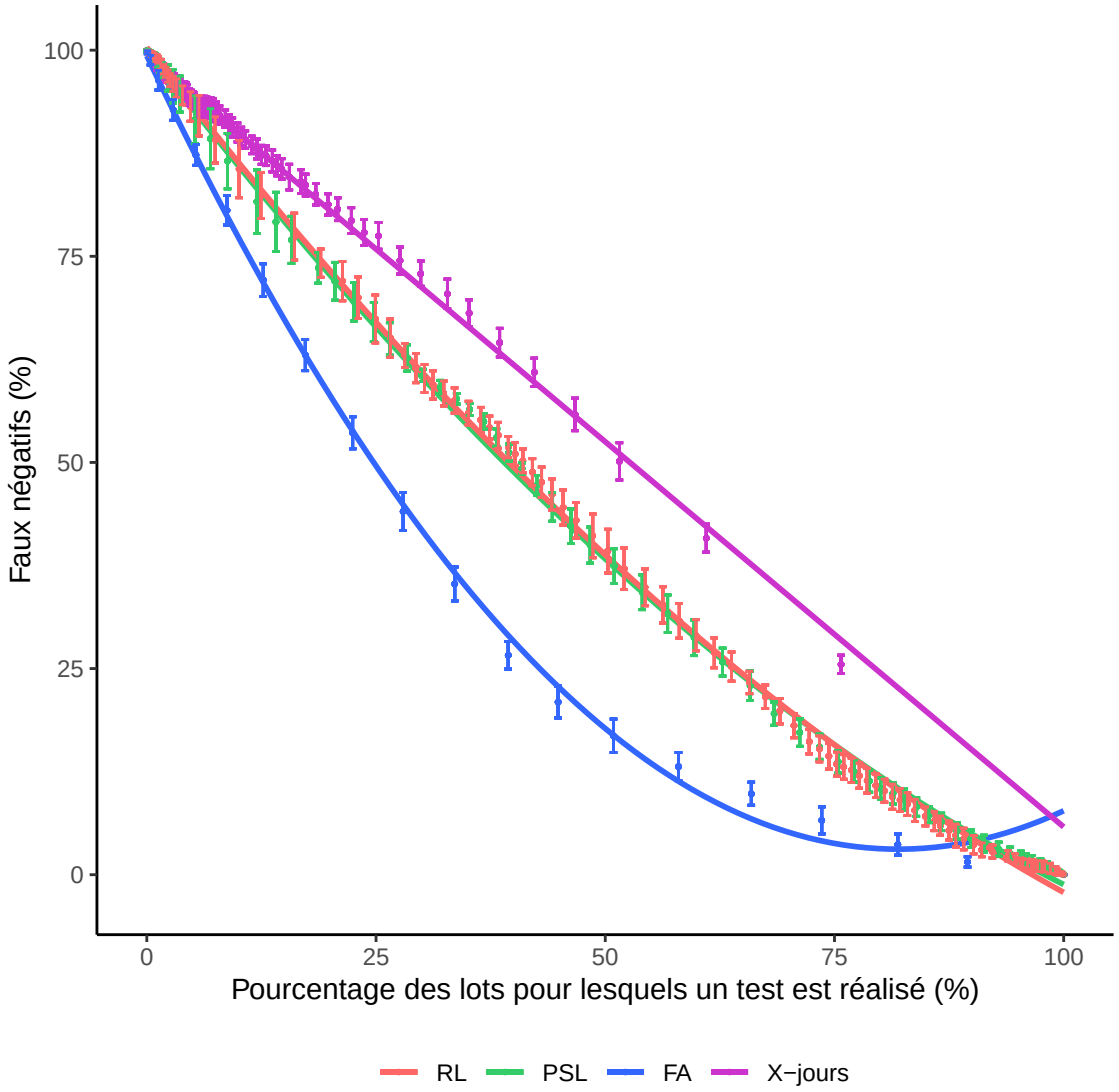

FIGURE 4.3 – Taux de faux négatif (FN) en fonction du pourcentage de pièces testées avant la production pour les trois modèles prédictifs ainsi que la *règle des X-jours* avec intervalle de confiance à 95%.

La figure 4.3 permet de comparer la performance des trois modèles prédictifs ainsi que la règle d'affaires (*règle des X-jours*). La mesure de performance utilisée est le taux de faux négatifs. En plus de permettre de comparer tous les modèles sur un même graphique, la figure 4.3 permet à l'entreprise d'établir un budget de test en fonction de la quantité de pièces potentiellement non conformes qu'elle est prête à accepter. Aussi, on peut constater sur la figure 4.3 que les modèles PSL et RL ont plus ou moins le même taux de faux négatifs à un pourcentage de tests égal. Le modèle qui performe le mieux, avec un taux de faux négatifs le plus bas, est le modèle de forêts aléatoires. Il est

important de noter que tous les modèles sont plus performants que la règle d'affaires établie préalablement par *Duvaltex*.

Dans le tableau 4.2, on compare le taux de faux négatifs obtenu par chacun des modèles pour le même taux de test (7%) que la règle initiale des 180 jours que Duvaltex s'était donnée afin de tester les pièces de tissu préalablement à la teinte. On constate que le modèle *FA* améliore d'environ 12% la détection des pièce problématiques versus la *règle des 180 jours*. Il est à noter que les modèles *RL* et *PSL* ont des performances pratiquement égales pour ce taux de test.

Modèles	Résultats
Régression logistique	89.1 ± 4.5
Règle des 180 jours	92.8
Forêt Aléatoire	80.8 ± 2.0
Projection sur les structures latentes	89.3 ± 5.9

Tableau 4.2 – Taux de faux négatif des modèles pour le même pourcentage de pièces de tissus testées que la *règle des 180-jours*

Cependant, puisque la relation entre le taux de faux négatifs et le pourcentage des lots testés avant la production n'est pas linéaire pour les trois modèles, l'écart entre le taux de faux négatifs de la *règle des X-jours* et ceux des autres modèles est plus important pour des valeurs plus élevées de pourcentage des lots testées. Par exemple, pour un pourcentage de tests avant la production de 25%, on remarque une diminution de 28% du taux de faux négatifs obtenus avec le modèle des *FA* versus ce qu'on obtiendrait pour le même pourcentage de tests avec la *règle des X-jours*.

4.3 Impact de la méthode du partitionnement du jeu de données

À la section 4.2, le jeu de données avait été séparé chronologiquement et les plus anciennes données étaient utilisées pour l'entraînement et la performance avait ensuite été mesurée sur les données les plus récentes du jeu de données. Dans cette section-ci, les trois modèles ont été entraînés et testés sur deux partitions du jeu de données initiales (tout en conservant les mêmes hyperparamètres). Tout en conservant le même pourcentage de données dans chacun des jeux de données (75% et 25%), les trois modèles ont

été entraînés et testés sur des jeux de données séparés aléatoirement, puis de manière antichronologique.

Pour le modèle *PSL*, on note à la figure 4.4 que le modèle performe mieux, pour un pourcentage de test inférieur à environ 70%, lorsque le partitionnement du jeu de données se fait en mode aléatoire et antichronologique que lorsqu'en mode chronologique. En haut d'un pourcentage d'environ 70% de tests, on constate qu'on obtient une performance similaire avec les trois méthodes de partitionnement.

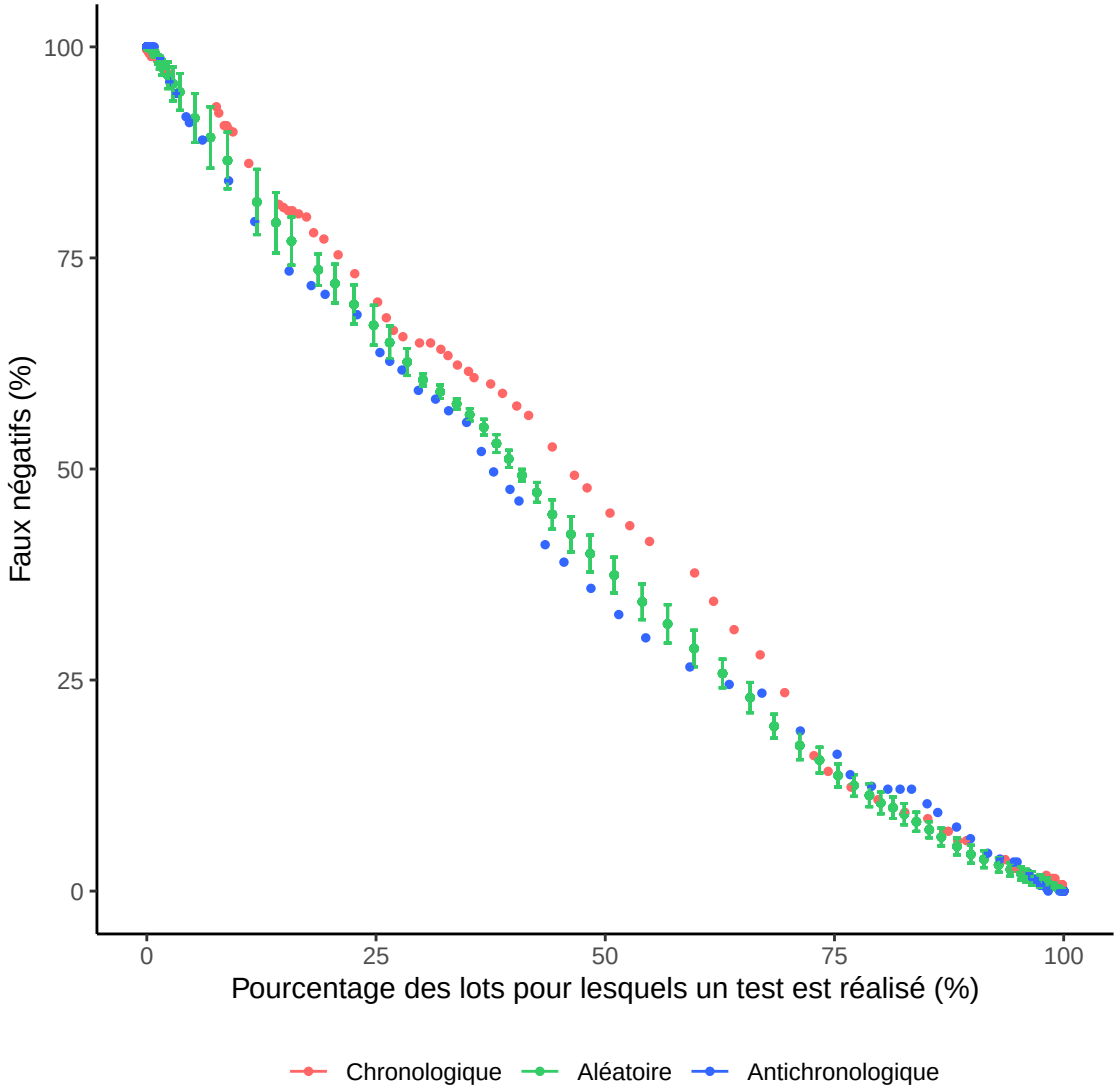

FIGURE 4.4 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *PSL* (Entraînement/Test : 75%/25%)

Pour le modèle *RL*, on constate à la figure 4.5 que les résultats obtenus pour les trois méthodes de partitionnement sont similaires à ceux obtenus avec le modèle *PSL*.

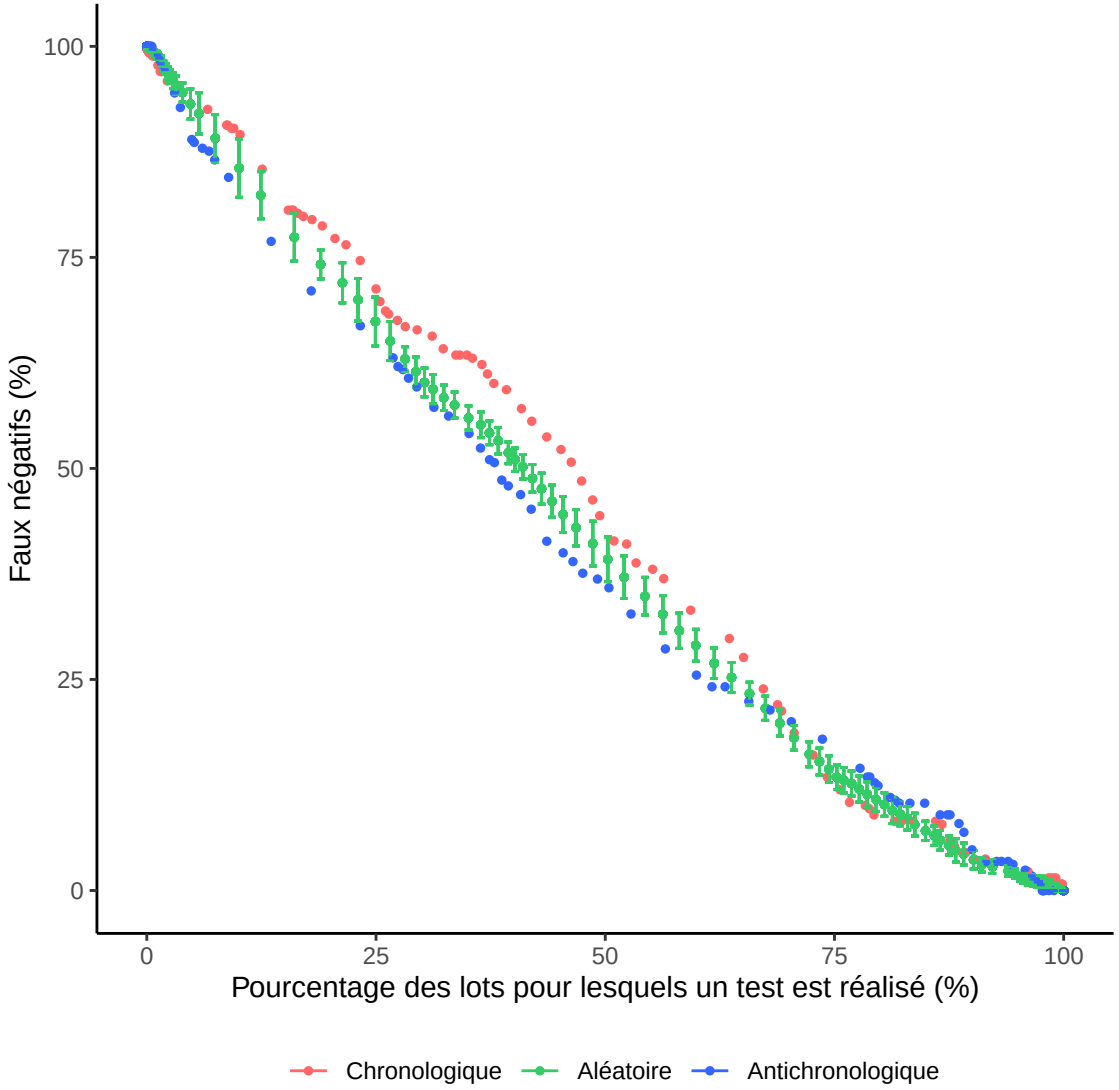

FIGURE 4.5 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *RL* (Entraînement/Test : 75%/25%)

Le modèle *FA* semble moins sensible aux changements de partitionnement que les deux autres modèles, comme on peut le voir à la figure 4.6. Toutefois, le partitionnement chronologique semble donner les pires résultats (le taux de faux négatifs le plus élevé) pour un pourcentage de tests jusqu'à environ 20% et d'environ 50% jusqu'à 100%. De 20 % à 50% environ, le mode de partitionnement qui donne le pire résultat avec ce

modèle est le mode antichronologique.

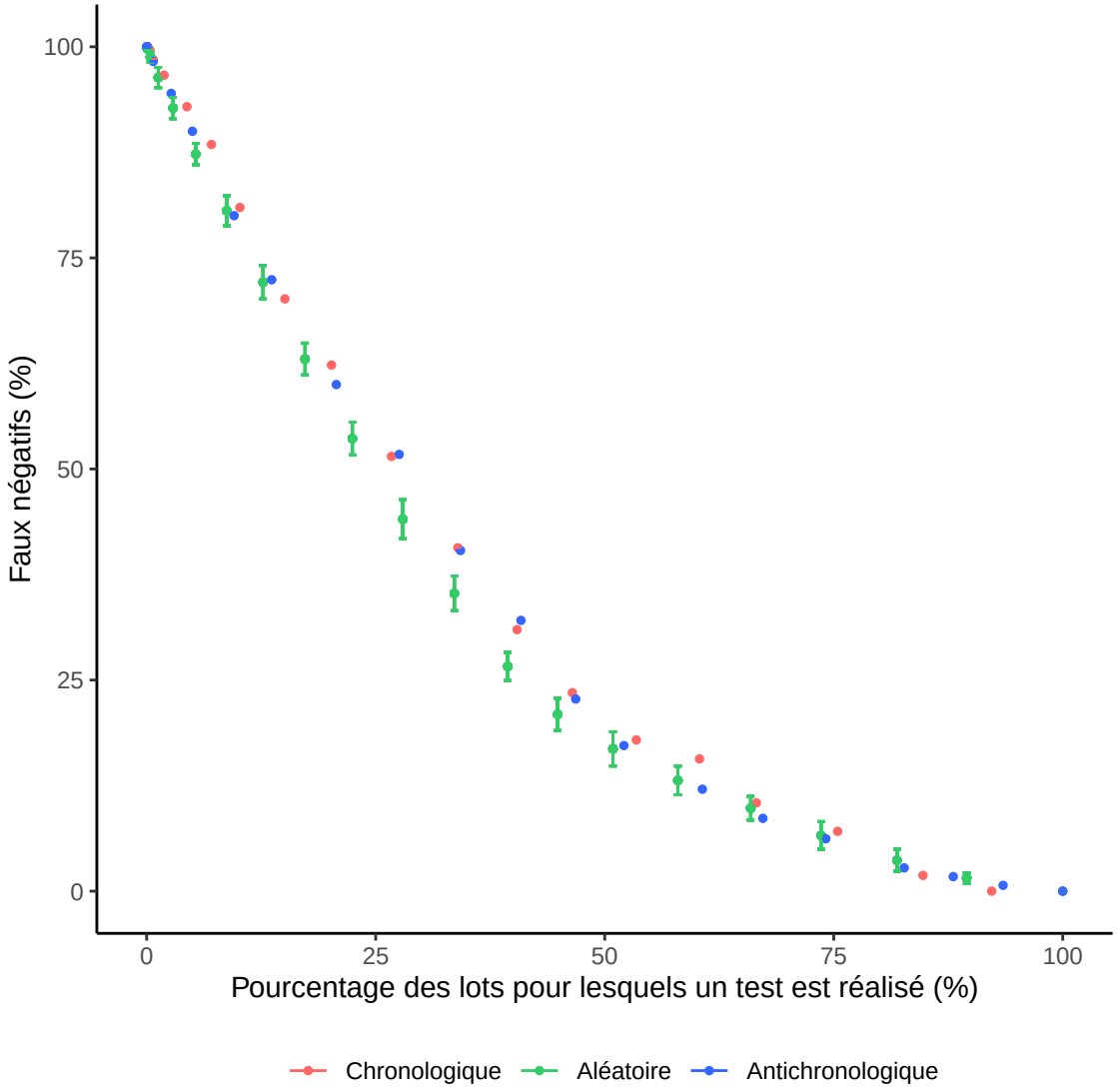

FIGURE 4.6 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents modes de partitionnement des données avec le modèle *FA* (Entraînement/Test : 75%/25%)

On peut supposer que les différences de performance selon le mode de partitionnement peuvent être dû au différent éventail de produits auquel le modèle a été exposé durant l'entraînement. On peut penser qu'un modèle exposé à plus de produits sera capable de mieux généraliser sur le jeu de données de test.

4.4 Impact de la taille du jeu de données d'entraînement

Dans cette section-ci, la taille du jeu d'entraînement et de test a été modifiée. La taille du jeu de données d'entraînement est passée de 75% de la taille du jeu de données initial à 60% et la taille du jeu de données de test est passée de 25% à 40%. La méthode de partitionnement du jeu de données est chronologique.

Pour le modèle *PSL*, on peut voir qu'une meilleure performance est obtenue avec une séparation entraînement/test de 60 %/40% pour des pourcentages de lots testés allant jusqu'à 65% environ. En haut de 65%, c'est la séparation entraînement/test 75%/25% qui obtient le taux de faux négatifs le plus bas. Il est important de noter que la différence de performance obtenue avec l'une ou l'autre des séparations n'est pas énorme.

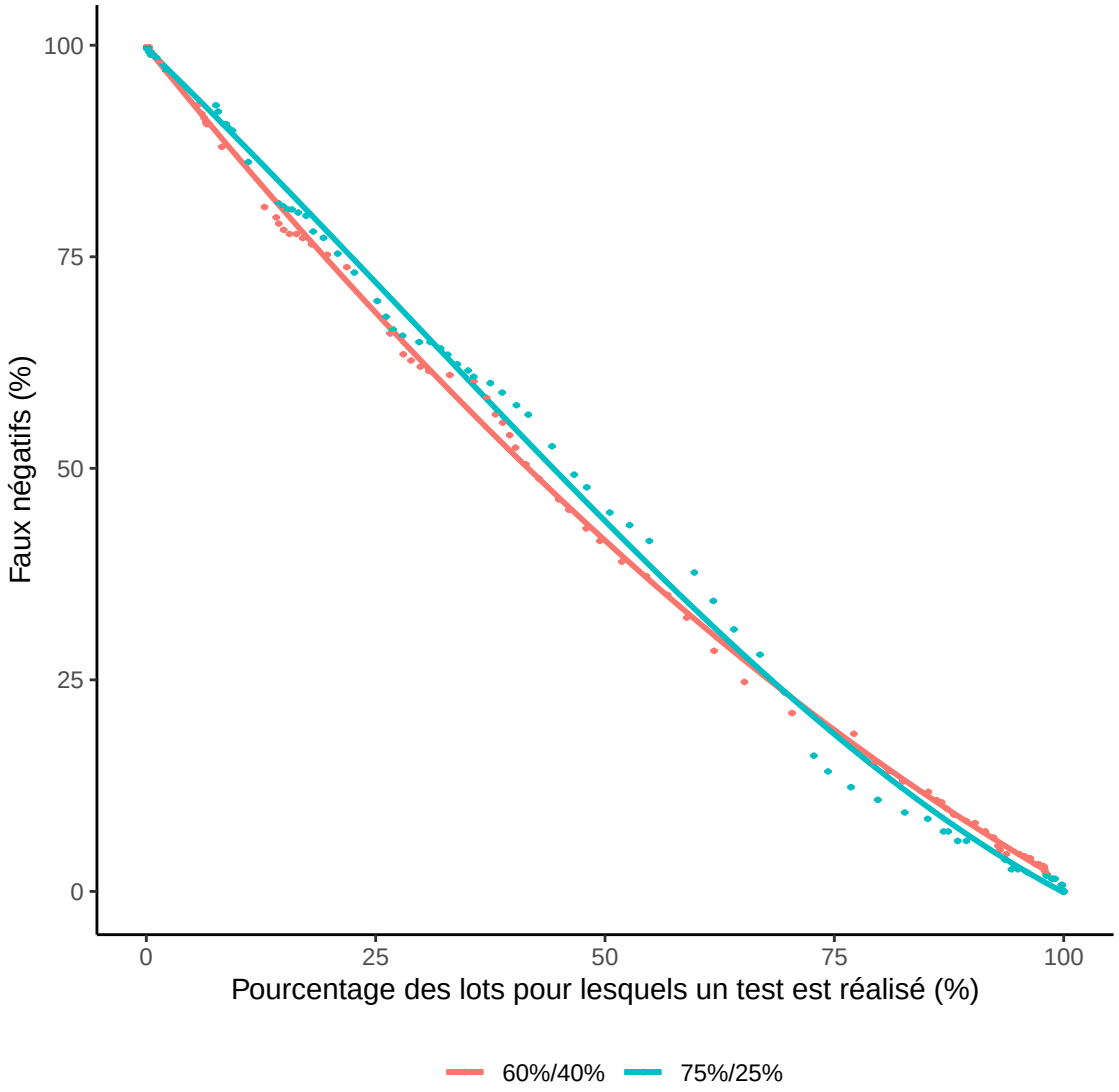

FIGURE 4.7 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentages de division du jeu d'entraînement/test avec le modèle *PLS*

La performance du modèle *LR* a été évaluée avec un partitionnement chronologique sur deux valeurs de séparation du jeu de données initial en jeu de données d'entraînement et de test. On peut voir qu'une meilleure performance est obtenue avec une séparation entraînement/test de 60 %/40% pour des pourcentages de lots testés allant jusqu'à 80% environ. Après 80%, la performance des deux séparations est plus ou moins égale.

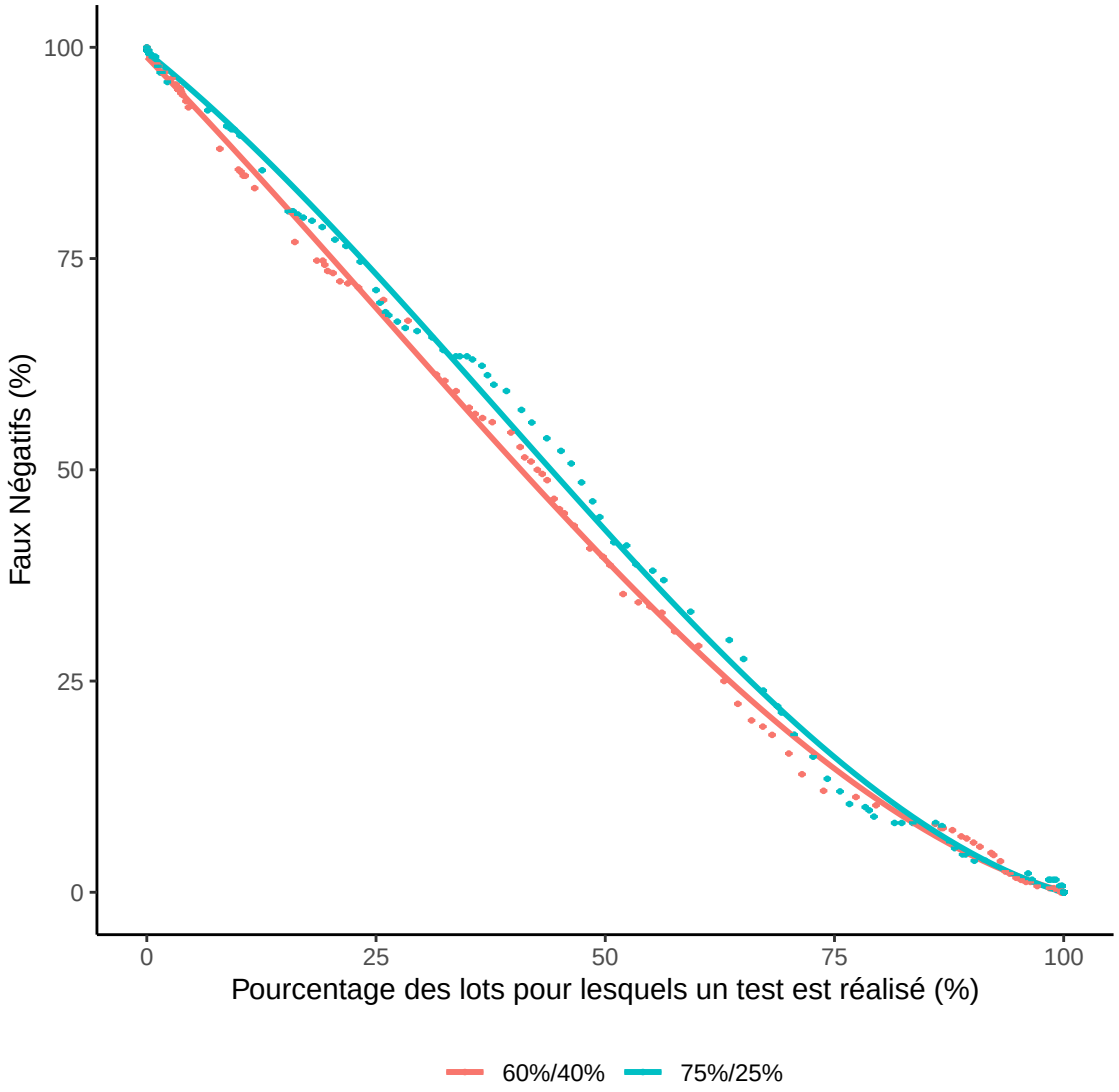

FIGURE 4.8 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentage de division du jeu d'entraînement/test avec le modèle *LR*

La performance du modèle *FA* a été évaluée avec un partitionnement chronologique sur deux valeurs de séparation du jeu de données initial en jeu de données d'entraînement et de test. On peut voir qu'une meilleure performance est obtenue avec une séparation entraînement/test de 60 %/40% pour des pourcentages de lots testés allant jusqu'à 75% environ. Après 75%, c'est la séparation entraînement/test 75%/25% qui obtient le taux de faux négatifs le plus bas. Il est important de noter que la différence de performance obtenue avec l'une ou l'autre des séparations pour ce modèle est moindre que les autres

modèles.

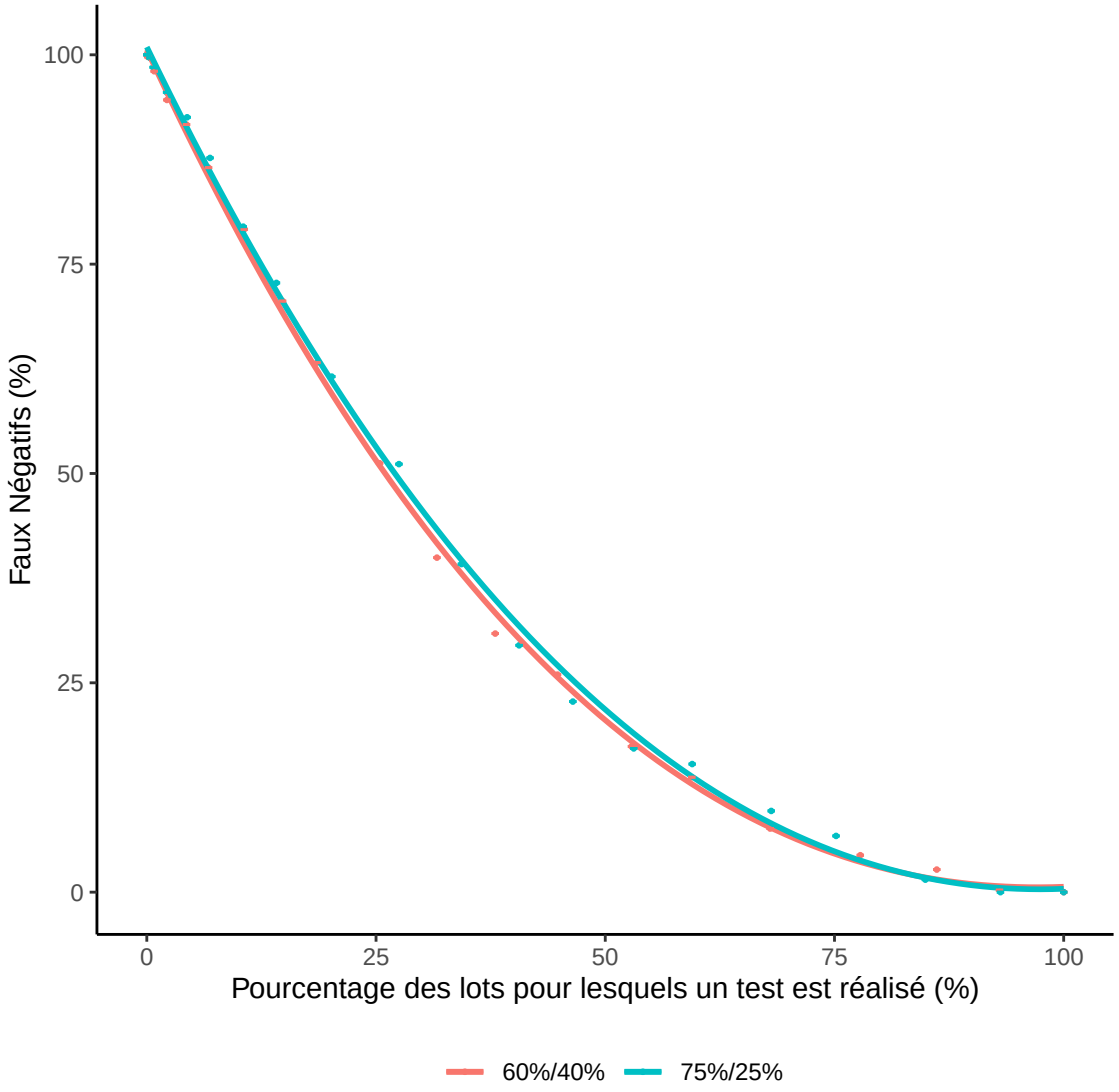

FIGURE 4.9 – Taux de faux négatifs (FN) en fonction du pourcentage de pièces testées avant la production pour les différents pourcentages de division du jeu d'entraînement/test avec le modèle *FA*

Les différences de performance des modèles peuvent indiquer que les valeurs des variables peuvent avoir fluctuées à travers le temps. Les résultats du mode de partitionnement chronologique ont été choisis afin de comparer les modèles, car il est celui qui fait le plus de sens si on implantait un modèle en production. Dans un contexte de production, on pourrait seulement s'entraîner sur des données historiques et on observerait la performance du modèle sur les données les plus récentes (en temps réel).

4.5 Importance des variables

Il est possible d'obtenir une idée de l'importance des variables dans le modèle *FA*. En effet, il a été établi qu'en permutant les valeurs d'une variable prédictive X, le lien de corrélation avec la variable réponse Y est rompu (Strobl et collab., 2008). Donc, si l'on permute les valeurs d'une variable qui est fortement corrélée à la variable réponse, on devrait voir une grande différence entre les valeurs prédites après la permutation et celles obtenues avant. On affirme alors que plus la différence de prédiction est grande, plus la variable est importante pour le modèle. Cependant, on ne peut pas parler de cause à effet, car les valeurs des données ne sont pas variées de façon indépendante et obtenues lors d'expériences contrôlées. On n'a pas non plus d'indications sur la façon dont la variable prédictive importante influence la variable qu'on tente de prédire. À la figure 4.10, on retrouve l'importance des variables utilisées dans le modèle *FA*.

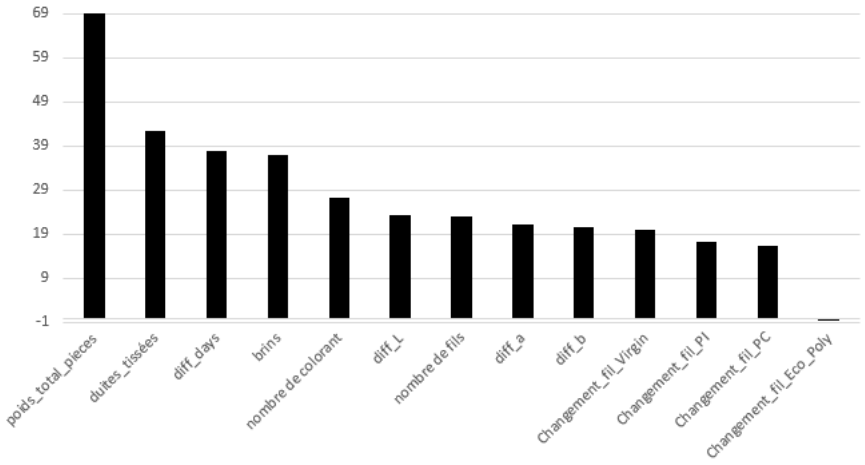

FIGURE 4.10 – Importance des variables pour le modèle *FA*

On peut voir sur la figure 4.10 que le nombre de jours depuis la dernière production (*diff_ days*) est parmi les trois variables les plus importantes pour le modèle *FA*. L'intuition de l'équipe de *Duvaltex* qu'il y avait peut-être une corrélation entre le nombre de jours écoulés depuis la dernière production d'un produit et la propension d'une pièce de tissu à être non-conforme n'est pas complètement fausse. Dans les variables les plus importantes, on voit aussi quelques caractéristiques du produit en tant que tel qui ne change pas à chaque production (*duites_ tissées* et *brins*). Cela pourrait indi-

quer que certains produits sont plus complexes à teindre que d'autres. Finalement, la variable qui semble avoir le plus d'importance pour le modèle est le poids de la pièce (*poids_total_pieces*). Lors de l'étude menée par Lajoie et collab. (2019), cette variable a également été identifiée comme importante.

Chapitre 5

Discussion

Dans la section précédente, la performance des trois modèles a été comparée. Ces trois modèles appartiennent à deux catégories : *FA* est un modèle non-linéaire tandis que *PSL* et la *RL* sont des modèles linéaires. On constate que le modèle non-linéaire, *FA*, performe mieux que les modèles linéaires. On peut supposer que les relations entre les variables qui régissent la conformité de la couleur sont complexes et plus facilement approximées par un modèle non-linéaire. De plus, la performance du modèle *FA* semble moins affectée par le changement de partition dans les données ce qui peut laisser entrevoir une plus grande généralisation du modèle que les autres.

L'utilisation du modèle *FA* présenté à la section précédente permet à Duvaltex d'établir une analyse financière de leur stratégie de test. Il serait possible d'établir le coût d'un faux négatif en regardant les coûts engendrés en moyenne par une pièce de tissu non conforme. Ces coûts peuvent être liés aux délais de livraison des commandes pour des clients (s'il y a des pénalités associées à celle-ci). Les pièces non conformes affectent aussi la productivité de l'usine. Même si elles peuvent être reteintes, ces pièces prennent le double de main-d'œuvre, temps machine et matières premières (colorants, produits chimiques etc.) pour être produit qu'une pièce conforme. De plus, si la couleur de la pièce de tissu ne peut pas être reprise pour atteindre la couleur désirée, la pièce devra être teinte dans une couleur plus foncée et neutre (noir, par exemple) et l'on devra trouver une façon de la vendre, probablement à un prix moindre que celui établi initialement. Ce scénario est doublement pénalisant, car, en plus d'avoir pris deux fois plus de ressources qu'une pièce régulière, cette pièce non conforme sera vendue moins cher.

Afin de compléter leur analyse financière, les coûts associés aux tests des lots avant la production devront être établis par l'équipe de Duvaltex. Afin d'effectuer ces tests, un morceau de tissu de la pièce tissée doit être prélevé et amener à l'équipe de coloristes afin de simuler sa teinte en laboratoire avec tous les produits chimiques et colorants qui entre en jeu dans la teinte régulière de cette pièce de tissu. Ensuite, la pièce doit être séchée et sa couleur analysée par un(e) coloriste. Il y a donc un certain coût de main-d'œuvre et de matière première associé à ces tests. De plus, ces tests prennent du temps à effectuer, donc la planification de la production devra s'ajuster en conséquence. Par exemple, on pourrait teindre des pièces à faible risque d'être non conforme (qui n'ont pas besoin d'être testée) à travers la teinte de pièces qui ont besoin d'être testées afin de maximiser l'occupation des machines et ne pas avoir des machines qui ne teignent pas de pièces, car elles sont en attente du résultat du test.

En fonction de cette analyse, la direction pourra décider quel pourcentage de test est optimal pour la compagnie. Chaque pourcentage de test est associé avec une probabilité que la pièce soit non conforme. Alors, on peut convertir ce pourcentage pour obtenir la probabilité qui sera le seuil de notification pour le modèle *FA*. L'équipe de coloristes sera notifiée dès qu'une pièce dépassera cette probabilité afin qu'elles soient testées avant de lancer la teinte de celle-ci en production.

Avant de mettre un modèle d'apprentissage automatique en production, il y a plusieurs points à considérer si l'on veut éviter certains problèmes de performance. Un des défis est d'implanter en ligne les transformations nécessaires aux multiples sources de données utilisées par le modèle (Schelter et collab., 2018). C'est l'une des étapes dont les efforts et le temps requis pour l'effectuer sont le plus souvent sous-estimés. Pour des fins de reproductibilité, il est recommandé de conserver les métadonnées des modèles implantés en production. Lorsqu'on parle de métadonnées, cela concerne plusieurs informations qui peuvent être utiles lorsqu'on veut améliorer et réentraîner le modèle. Des informations comme qui a créé le modèle initialement, quelles données ont été utilisées pour l'entraînement et l'évaluation du modèle ainsi que les transformations qui ont été appliquées sur ces données (Schelter et collab., 2018). Le but de garder ces informations est, entre autres, de pouvoir comparer la performance de nouveaux modèles avec les anciens.

Garder ces métadonnées peut aussi servir lorsqu'on veut valider le modèle lorsqu'il y

a des changements dans les données. Il est aussi important que la façon de mesurer la performance des modèles reste la même. Même si l'on cherche à toujours améliorer un modèle, il est rare que l'on améliore drastiquement la performance sans demander une plus grande capacité de calcul, ralentir l'exécution du programme qui fournit la prédiction et même parfois demander plus de données (Schelter et collab., 2018). Les modèles d'apprentissage automatique devraient toujours être déployés en arrière-plan pour une certaine période afin de permettre aux gens de s'habituer à son utilisation et comparer les prédictions obtenues avec ce que l'on voit réellement sur le terrain.

Un autre point important lorsqu'un modèle est en production est de pouvoir détecter lorsqu'il doit être réentraîné. Cela doit se produire lorsque des changements sont effectués et que cela impacte les données sur lesquelles le modèle a été entraîné (Schelter et collab., 2018). Dans le cas spécifique de Duvaltex, cela pourrait être à une période fixe, hebdomadairement ou journalièrement par exemple. Il ne s'agit pas seulement de réentraîner le modèle lorsque de nouvelles variables sont disponibles. Par exemple, si un nouveau style de tissu est introduit en production et a des proportions de types de fils qui le composent complètement différentes que les produits dans les données sur lesquelles le modèle a été entraîné, il pourrait s'avérer judicieux de réentraîner le modèle dans une telle situation. Comme on fait l'entretien des capteurs physiques (recalibration périodique), les modèles prédictifs ont eux aussi besoin d'une certaine maintenance. En plus des transformations des données de multiples sources, on doit donc s'assurer que les distributions et valeurs de celles-ci ne s'éloignent pas trop des données initiales. Sinon, on doit réentraîner le modèle hors-ligne et ensuite le remettre en production. Plus ce processus sera robuste, plus on évitera que les performances du modèle se détériorent au fil du temps.

Conclusion

Ce mémoire aborde la problématique du contrôle qualité au sein d'une entreprise manufacturière, *Duvaltex*. Le but est de capturer les pièces de tissu susceptibles d'avoir une non-conformité de la couleur avant la teinte de celle-ci. À la suite d'une revue de la littérature, trois modèles ont été développés et comparés à une règle d'affaire qui se basait sur le nombre de jours (180 jours) depuis la dernière production afin de cibler les lots à risque.

Une analyse des systèmes utilisés et données générées par l'entreprise dans la cadre de la teinte de tissus a été effectuée. Cette analyse a permis d'extraire les données nécessaires au développement de modèles prédictifs. De nouvelles variables ont aussi été créées à partir des données brutes afin d'alimenter ces modèles.

Les modèles prédictifs proposés ont des taux de faux négatifs variant de 80,8% à 89,3%. Le meilleur modèle permet ainsi d'améliorer de 12 % la détection des pièces problématiques versus la règle d'affaire utilisée par l'entreprise. En plus de cette amélioration, le modèle proposé permet à l'entreprise d'établir une stratégie de tests en fonction du coût de ces tests et du coût de ne pas détecter des pièces de tissus problématiques.

Pour la suite des choses, comme les modèles développés n'ont pas été implantés en entreprise, une implantation pourrait faire partie des prochaines étapes. Cela pourrait permettre de peaufiner ceux-ci en fonction de leur performance. Aussi, d'autres sources de données externes à l'entreprise pourraient être ajoutées au modèle, comme l'analyse de qualité des fils lors de leur fabrication.

Bibliographie

Behera, B. et B. Karthikeyan. 2006, «Artificial neural network-embedded expert system for the design of canopy fabrics», *Journal of industrial textiles*, vol. 36, n° 2, p. 111–123.

Bishop, J., M. Bushnell et S. Westland. 1991, «Application of neural networks to computer recipe prediction», *Color Research & Application*, vol. 16, n° 1, p. 3–9.

Breiman, L. 2001, «Random forests», *Machine learning*, vol. 45, n° 1, p. 5–32.

Cadavid, J. P. U., S. Lamouri, B. Grabot et A. Fortin. 2019, «Machine learning in production planning and control : A review of empirical literature», *IFAC-PapersOnLine*, vol. 52, n° 13, p. 385–390.

Chaouch, S., A. Moussa, I. Ben Marzoug et N. Ladhari. 2019, «Colour recipe prediction using ant colony algorithm : principle of resolution and analysis of performances», *Coloration Technology*, vol. 135, n° 5, p. 349–360.

Chen, Z., C. Zhou, Y. Zhou, L. Zhu, T. Lu et G. Liu. 2018, «Multi-dimensional regression for colour prediction in pad dyeing», dans *International Conference on Cloud Computing and Security*, Springer, p. 675–687.

Clark, M. 2011, *Handbook of textile and industrial dyeing : principles, processes and types of dyes*, Elsevier.

Eldessouki, M., M. Hassan, K. Qashqary et E. Shady. 2014, «Application of principal component analysis to boost the performance of an automated fabric fault detector and classifier», *Fibres & Textiles in Eastern Europe*.

Fernaeus, Y., M. Jonsson et J. Tholander. 2012, «Revisiting the jacquard loom : Threads of history and current patterns in hci», dans *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, Association for Computing Machinery, New York, NY, USA, ISBN 9781450310154, p. 1593–1602, doi :10.

1145/2207676.2208280. URL https://doi-org.acces.bibl.ulaval.ca/10.1145/2207676.2208280.

Hasanbeigi, A. et collab.. 2010, «Energy-efficiency improvement opportunities for the textile industry», cahier de recherche, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Hastie, T., R. Tibshirani et J. Friedman. 2001, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Kandi, S. G. 2007, «Color recipe prediction by genetic algorithm», *Dyes and Pigments*, vol. 74, nº 3, p. 677–683.

Lajoie, P., J. Gaudreault, N. Lehoux et M. B. Ali. 2019, «A data-driven framework to deal with intrinsic variability of industrial processes : An application in the textile industry», *IFAC-PapersOnLine*, vol. 52, nº 13, p. 731–736.

Lee, C.-Y., J.-Y. Lin et R.-I. Chang. 2018, «Improve quality and efficiency of textile process using data-driven machine learning in industry 4.0», dans *International Symposium on Theory and Practice in IT, Engineering & Applied Sciences (TPIEA)*.

Lu, Y. 2017, «Industry 4.0 : A survey on technologies, applications and open research issues», *Journal of industrial information integration*, vol. 6, p. 1–10.

Metz, C. E. 1978, «Basic principles of roc analysis», dans *Seminars in nuclear medicine*, vol. 8, Elsevier, p. 283–298.

NPTEL. 2014, «Introduction to fabric manufacturing», URL https://nptel.ac.in/courses/116/102/116102005/.

Samuel, A. L. 1959, «Some studies in machine learning using the game of checkers», *IBM Journal of research and development*, vol. 3, nº 3, p. 210–229.

Schelter, S., F. Biessmann, T. Januschowski, D. Salinas, S. Seufert et G. Szarvas. 2018, «On challenges in machine learning model management», .

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin et A. Zeileis. 2008, «Conditional variable importance for random forests», *BMC bioinformatics*, vol. 9, nº 1, p. 1–11.

Su, T.-L. et C.-F. Lu. 2011, «Automated vision system for recognising lycra spandex defects», *Fibres & Textiles in Eastern Europe*, vol. 19, nº 1, p. 43–46.

Tao, F., Q. Qi, A. Liu et A. Kusiak. 2018, «Data-driven smart manufacturing», *Journal of Manufacturing Systems*, vol. 48, p. 157–169.

Weatherall, I. L. et B. D. Coombs. 1992, «Skin color measurements in terms of cielab color space values», *Journal of investigative dermatology*, vol. 99, n° 4, p. 468–473.

Wuest, T., C. Irgens et K.-D. Thoben. 2014, «An approach to monitoring quality in manufacturing using supervised machine learning on product state data», *Journal of Intelligent Manufacturing*, vol. 25, n° 5, p. 1167–1180.

Wuest, T., D. Weimer, C. Irgens et K.-D. Thoben. 2016, «Machine learning in manufacturing : advantages, challenges, and applications», *Production & Manufacturing Research*, vol. 4, n° 1, p. 23–45.

Yildirim, P., D. Birant et T. Alpyildiz. 2018, «Data mining and machine learning in textile industry», *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, vol. 8, n° 1, p. e1228.

Yıldız, K., A. Buldu et M. Demetgul. 2016, «A thermal-based defect classification method in textile fabrics with k-nearest neighbor algorithm», *Journal of Industrial Textiles*, vol. 45, n° 5, p. 780–795.

Zhang, J. et C. Yang. 2014, «Evaluation model of color difference for dyed fabrics based on the support vector machine», *Textile Research Journal*, vol. 84, n° 20, p. 2184–2197.

Annexe A

Article de conférence (INCOM 2021 : 17th IFAC Symposium on Information Control Problems in Manufacturing)

Predictive Models to Improve the Quality of a Textile Dyeing Process

Catherine Bourdeau-Laferrière, Jonathan Gaudreault, Carl Duchesne

CRISI Research Consortium for Industry 4.0 Systems Engineering,
Université Laval, Québec City, Québec, CANADA
(e-mail:catherine.bourdeau-laferriere.1@ulaval.ca,
jonathan.gaudreault@ift.ulaval.ca, carl.duchesne@gch.ulaval.ca

Abstract: Industrial manufacturers are facing increased demand for custom products, which comes with an increase of process variability. For example, in the case study presented in this paper, the company put business rules in place to mitigate the non-conformity of color of the textiles they produce. We decided to use historical data to build predictive models to help the manufacturer design a more efficient quality control strategy. The performance of these models was compared to the business rule previously used by the company. It was found that the random forest model outperformed their business rule by 12% (reduction of false negative ratio) for the same number of quality tests performed. Also, the proposed method allows the business to better choose the tests to perform based on their budget and the number of non-compliant products that are deemed acceptable for them.

Keywords: Probabilistic & statistical models in industrial plant control, Industrial and applied mathematics for production, Industry 4.0, Data analytics, Data mining and multivariate statistics

1. INTRODUCTION

Today's challenge for industrial manufacturers is the fact that they do not produce a large volume of the same product but a small amount of a large variety of products. In this context, it is challenging to control the variability of the characteristics of the final product. For example, in the food processing industry the taste of cookies must be the same even if the flour comes from different crops of wheat. Variability can be introduced by the manufacturing process, raw materials, manual operations or other processes prior or parallel to the one of interest. It can be too costly or impossible to control all sources of variability. Therefore, the goal is to understand its impact on the process and learn to adapt it in order for the products to remain within the limit of conformity.

The case study presented here concerns a textile factory producing dyed textile for the office furnishings industry. For this specific market, the exact replication of the color from one batch to another is crucial. If the textile is not the exact same color (i.e., is "off-shade") the customer can return the order. This causes additional costs for the factory as they need to produce the textile again or re-dye it in some cases. In order to achieve this goal, the factory tests the resulting color on a small sample before producing a batch of a product that has not been manufactured in a long period of time (typically 180 days). However, we asked ourselves if by using historical data, we would be able to create a machine learning model which could predict if the batch is at risk of being off-shade, thus determining the relevance of testing this batch. This could reduce the number of tests or keep the same number of tests but better target which batch to test. By reducing the number of test, the factory would reduce labor costs while better targeting which batch to test would reduce the cost of non-quality.

The remainder of the paper is divided into four sections. Section 2 will present the textile production domain and the problem encountered by our research sponsor. Section 3 will detail the current business rule used by the manufacturer. Section 4 describes the proposed predictive models. Section 5 reports on the performance of these models. Section 6 concludes this article.

2. PRELIMINARY CONCEPTS

There are three main steps that are part of the process to manufacture the end product. The first step is to weave the yarns to make a fabric. The yarn used in this process can be made of different types of fibers depending on the product. The thread that lies on the length of the piece of fabric is called a *warp*. They are combined on a loom. The threads that are incorporated widthwise are called *wefts*. The final pattern of the fabric depends on the way warps and wefts are combined. At this point in the process, many quality issues can be encountered. Quality inspection can be labor and time consuming steps. Using machine learning for quality prediction can reduce production lead-time, improve client relationships and help to identify potential causes for quality defects (Krauß et al., 2020). However, Ogulata et al. (2006) show that some tests can be replaced by predictive models. They demonstrated that they could predict the elongation and recovery value of a fabric using neural networks and linear regression. Principal component analysis combined with a neural network was used by

Eldessouki et al. (2014) to detect different types of weaving defect based on fabrics images.

Once the fabric is made, it is then dyed. The dyeing process usually involves chemicals to clean the fabric and then dyes to obtain the desired color. After this step, it is dried before being inspected for quality defects. One of the most important goals in manufacturing fabric for furniture is obtaining the exact shade of color for each lot (Chen et al., 2018). The margin of error for the resulting color is very tight and considering the large number of different products being made, it is challenging to maintain a high operational efficiency. Many variables make the replication of the exact same color from batch to batch hard to achieve. The manufacturer needs to cope with a lot of off-shade batches that, unfortunately, are only detected at the end of the dyeing process at the quality inspection. Chen et al. (2018) demonstrated that they could predict the resulting color of the fabric based on the quantity of dye used and the properties of the fabric. The performance of support vector machine was compared to a neural network model. Moreover, Bishop et al. (1991) used a neural network model to define the concentration of dye to be used in order to obtain specific colors on nylon fabric.

3. 180 DAY VS X-DAY RULE

For Duvaltex, obtaining the right colour every time they dye a fabric is critical. In previous work by Lajoie et al. (2019), a correlation was established between many measurable variables on the dyeing process (amount of water used, fabric rotation speed, fabric weight, dye characteristic etc.) and the compliance of the resulting color to the standard. Following this work, many corrective actions were put in place at the factory in order to reduce the number of off-shade pieces of fabric.

However, even after applying these new control methods, non-compliant fabric are still observed. Therefore, with the aim of identifying production runs with a high risk of resulting in not-compliant textile color, Duvaltex established the so-called *180-day rule*. For each batch of a certain product that has not been dyed for the last 180 days, they assume that some parameters might have changed since the last production (yarn batch, dyes batch, etc.). Then, they do a pre-dye test which consists of dyeing a small piece of the textile in the laboratory and adjusting the parameters of the process if needed.

The capacity to detect problematic batches using the rule currently used by our partner was assessed. By using the historical data, we were able to confirm if the batches that were last produced for more than 180 days were, at the end of the process, non-compliant. Moreover, the rule was then generalized to become the *X-day rule* where time intervals other than 180 days (from 1 to 500 days) were tested.

In order to compare the performance of the *180-day rule* and the *X-day rule*, we chose three metrics: *False Negative ratio* (FN), *False Positive ratio* (FP) and *Accuracy* (Acc). False negatives are the number of batches that would not have been tested using the *X-day rule* but that were not compliant at the end of the process. *False negatives* represent the number of problematic pieces for which a pre-dye test would not have been performed. The false negative ratio is the number of false negatives over the total number of non-compliant batches. False positives are the number of batches that would have been tested using the *X-day rule* but that were compliant at the end of the process. *False positives* can be viewed as the number of useless pre-dye tests because the pieces would have respected the colour conformity threshold. The false positive ratio is the number of false positives over the total number of compliant batches. *Accuracy* (Acc) is defined by the ratio of correctly classified products over the total number of observations.

To be able to compare the performance of the *180-day rule* and the *X-day rule*, the same dataset was used. For the "180-day rule", we obtain an accuracy of 73.2%, a false negative ratio of 92.8% and a false positive ratio of 6.6% were obtained. Figure 1 shows the performance of the "X-day rule". It is easily noticed that as FN increases FP decreases. The goal is therefore to select the number of days resulting in the smallest FP and FN possible. For our research sponsor, Duvaltex, the most important metric is reducing the number of FN.

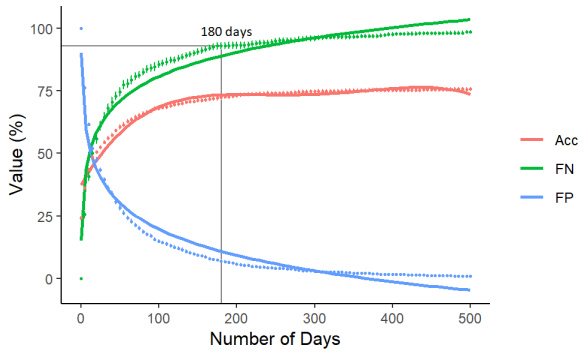

Fig. 1. Accuracy (Acc), False Negative (FN) and (FP) according to the minimal number of days between two batches of the same product that triggers a pre-test.

It was decided to put the number of pre-dye tests on the x-axis so the factory can see more easily what the difference would be in the number of non-compliant fabrics as they increase or decrease the number of days (and thus, as they decrease or increase the number of tests that are performed) in the *X-day rule*. Figure 2 shows how the FN ratio decreases as the percentage of total production requiring a pre-dye test increases. If we test 100% of the pieces that go into production, FN would be at 0%, but it is not feasible from a labor point of view to test every piece of textile before dyeing it. Indeed, the figure enables the company to establish their budget for pre-dye testing as a tradeoff between the cost of pre-dye testing (material, labour time, etc.), and the cost of non-compliant pieces of textile. As one can see, the relation between the number of tests and the number of days since the last production is linear, which is not really good. FN will decrease linearly as the number of tests performed increases.

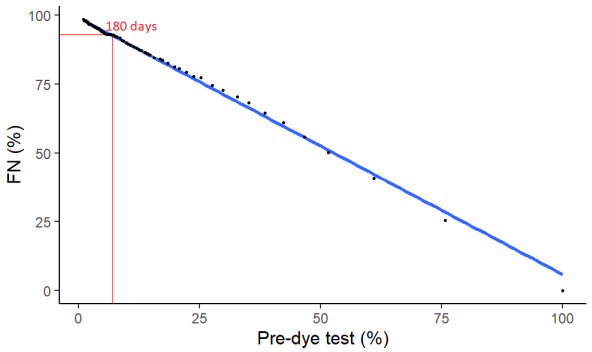

Fig. 2. X-day rule False Negative (FN) ratio according to number of pre-test.

4. PREDICTIVE MODELS

The objective of these experiments is to compare the performance of the *X-day rule* to predictive models. The goal is to provide a method enabling the factory to determine the number of tests and the amount of non-compliant fabric yielding the lowest cost. We will use historical data to build three predictive models: random forest, partial least squares and a logistic regression.

4.1 Data

To build the three models, data on product style, the weaving process and the dyeing raw material was used. We had access to historical data containing information about 4302 production batches of 693 different products. The different variables used in the models are listed in Table 1. The first category (1) of variables describes the characteristics of each product. The values of these variables remain the same for a given product and do not change for each production run.

For the second category (2) of variables, values may change from batch-to-batch of the same product. There are four types of threads that are used by our partner: Eco Polyester, Post Industrial Recycled, Post Consumer Recycled and Virgin. Our partner noticed that there is a higher variation in the thread's color from one batch to another when the type of thread is Post Consumer Recycled. Therefore, it was decided to include this information in the models.

The dyed batch variables (3) describes the color change of each dye and the number of days since the last production of the same product. The changes in dye color are quantified by a few color features. The color of each dye is described by three values: lightness, red-green value and blue-yellow value. The lightness of a dye measures the darkness of its color. The red-green value measures how much the color tends toward green or red. The blue-yellow value measures how much the color tends toward one or the other color on the blue-yellow axis. Before using the dyes in production, the color of each batch of dyes is measured using the LAB color space.

4.2 Partial Least Squares

The first model we developed uses Partial Least Squares (PLS). This method, introduced by Wold (1983) is an approach that combines concepts from Principal Component Analysis (PCA) and multiple linear regression (Abdi, 2010). Partial least squares deals effectively with large datasets of highly correlated variables. In previous studies, PCA was used to reduce dimensionality prior to using another predictive model to detect textile defects such as color differences on dyed fabric (Zhang and Yang, 2014) and fabric pilling (Jing et al., 2012). In our experimentations, the PLS hyperparameter is the number of principal components and it was selected equal to 10 using the training dataset.

4.3 Random Forest

This method was introduced by Breiman (2001). Random Forest (RF) is an ensemble method that uses decision trees as weak learners. Decision trees are well known for their easy interpretation but they often overfit the data. To prevent this, RF uses a large number of decision trees each grown over a bootstrap sample of the initial data with a randomized sample of the initial variables (Verikas et al., 2011). In the survey conducted by Verikas et al. (2011), it was reported that RF was used for a number of applications ranging from natural language processing to classification of image data. In the manufacturing field, RF was used by Puggini et al. (2015) to detect defects on semiconductor using manufacturing process data. This technique was also used in a study conducted by Wu et al. (2017) for preventive maintenance of manufacturing tools. In our experimentations, the hyperparameters for this model are the number of trees (set to 500), the number of variables to use to make the split at each node (set to 3), and the minimal number of observations that can be in a terminal node (set to 4). Many hyperparameter values were tried and the values were selected based on the best performance on the training dataset.A validation set was

Table 1: List of textile products, weaving batch and dye batch characteristics variables with their corresponding descriptions

Category	Name	Description
Textile product characteristics(1)	Number of weft	Number of weft use in the weaving process
	Number of warp	Number of warp use in the weaving process
	Number of thread	Number of thread use in the weaving process
	Number of dyes	Number of different dye in the recipe of the product
Weaving batch characteristics(2)	Eco Polyester Thread Changes	Ratio of the amount of thread that is not from the same batch as the ones that were used in the prior production for the same product
	Post Industrial Recycled Thread Changes	
	Post Consumer Recycled Thread Changes	
	Virgin Thread Changes	
	Weight of textile piece	The weight of the piece before dyeing process
Dye batch characteristics(3)	Dye Color Changes on Illuminant Axis	The ratio of changes on the three different measures of the dyes colours that have changed since the last production of the same product
	Dye Color Changes on Yellow-Blue Axis	
	Dye Color Changes on Red-Green Axis	
	Number of days	Number of days since the last production of the same product

not created to adjust the hyperparameters because of the small amount of data available to fit the model.

4.4 Logistic Regression

Since the 1960s, this technique has been used in multiple fields as a predictive model when the desired output is a categorical variable (Candanedo et al., 2018). It is similar to a linear regression. However, instead of predicting one value, it can return multiple values which represent the probability of the observation to be in a certain category. In the manufacturing industry, it was used to build a model for the preventive maintenance of Heating, Ventilation and Air Conditioning Systems (Candanedo et al., 2018).

5. EXPERIMENTS

From the databases of our industrial partner, fifteen months of weaving and dyeing process data was extracted. Only the information about the textile pieces that were dyed was retained. The data was split randomly in order to obtain training (75% of the data) and testing (25% of the data) datasets.

Three predictive models were built using the variables previously described: RF, PLS and a logistic regression model. The performances of these models were compared to the "X-day rule" previously used by our industrial partner. Indeed, each model returns a continuous value between 0.0 and 1.0 for each batch in the test dataset. It represents the probability of a batch to be non compliant. A threshold (between 0.0 and 1.0) is then selected to classify the batch as good quality or non-compliant. The lower the threshold the higher the number of batches that will need to be tested by the laboratory technicians. The percentage of tests shown on the x-axis of Figure 3 is calculated by varying the threshold for which we consider a batch to be predicted as problematic.

Figure 3 shows the FN ratio for all the models developed in this study and the X-day rule previously used by the manufacturer. While allowing the comparison of all the models together, this chart also provides the manufacturer with a tool to establish a pre-dye testing budget by balancing the number of tests to perform and number of false negatives allowed. The PLS and logistic regression models give both more or less the same performance. Moreover, they both out performs the "X-day rule" but their performance is lower than the RF models. In addition, the RF model has the lowest FN ratio compared with the other models including the "X-day rule".

Compared with the initial "180-day rule", the RF model FN ratio is 12 % lower for the same number of pre-dye tests. Thus, by replacing this rule by the RF model, Duvaltex could expect a reduction in the number of problematic batches not tested before production by 12%. On the other hand, if Duvaltex decided that they want to perform a greater number of pre-dye tests, they could reduce even further the number of non-compliant batches. For example, if they decided to test approximately 30% of the products before the dyeing process, selecting them using the RF model, they would be able to correct 35.6% ±2.2% more problematic product recipes than if they were relying on testing on the number of days since the last time the product was dyed. Over a period of approximately 15 months, this could prevent the non-conformity of an additional 396 batches if this were to generate economic benefits.

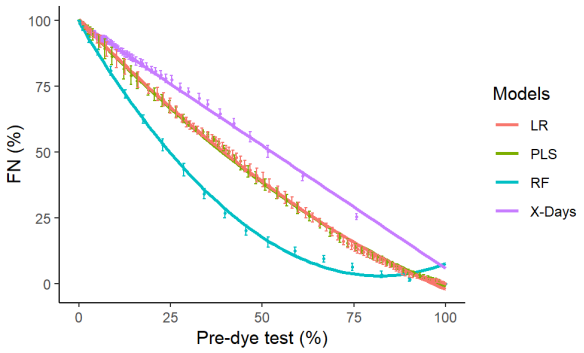

Fig. 3. LR, PLS RF and X-Days rule False Negative (FN) ratio according to number of pre-tests. Results reported for 10 replications and a 95 % confidence interval

6. CONCLUSION

We compared three models based on production data for predicting quality of a product in the textile industry with the current best practice ("180-days rule") used at our partner site. It shows that using data mining techniques trained on data identified by domain experts yields better risk assessment of product quality than a simple business rule. Using the random forest models could lower the number of fabrics showing color non-compliance defects. Finally, it allows the business to improve the efficiency of their pre-dye testing by understanding how many non-compliant fabrics they can prevent by increasing the amount of tests performed by their laboratory technician team. It enables the company to establish their budget for pre-dye testing as a tradeoff between the cost of pre-dye testing and the cost of non-compliant pieces of textiles.

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1), 97–106.

Bishop, J., Bushnell, M., and Westland, S. (1991). Application of neural networks to computer recipe prediction. *Color Research & Application*, 16(1), 3–9.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Candanedo, I.S., Nieves, E.H., González, S.R., Martín, M.T.S., and Briones, A.G. (2018). Machine learning predictive model for industry 4.0. In L. Uden, B. Hadzima, and I.H. Ting (eds.), *Knowledge Management in Organizations*, 501–510. Springer International Publishing, Cham.

Chen, Z., Zhou, C., Zhou, Y., Zhu, L., Lu, T., and Liu, G. (2018). Multi-dimensional regression for colour

prediction in pad dyeing. In *International Conference on Cloud Computing and Security*, 675–687. Springer.

Eldessouki, M., Hassan, M., Qashqary, K., and Shady, E. (2014). Application of principal component analysis to boost the performance of an automated fabric fault detector and classifier. *Fibres & Textiles in Eastern Europe*.

Jing, J., Zhang, Z., Kang, X., and Jia, J. (2012). Objective evaluation of fabric pilling based on wavelet transform and the local binary pattern. *Textile Research Journal*, 82(18), 1880–1887.

Krauß, J., Pacheco, B.M., Zang, H.M., and Schmitt, R.H. (2020). Automated machine learning for predictive quality in production. *Procedia CIRP*, 93, 443–448.

Lajoie, P., Gaudreault, J., Lehoux, N., and Ali, M.B. (2019). A data-driven framework to deal with intrinsic variability of industrial processes: An application in the textile industry. *IFAC-PapersOnLine*, 52(13), 731–736.

Ogulata, S.N., Sahin, C., Ogulata, R.T., and Balci, O. (2006). The prediction of elongation and recovery of woven bi-stretch fabric using artificial neural network and linear regression models. *Fibres & Textiles in Eastern Europe*, 14(2), 56.

Puggini, L., Doyle, J., and McLoone, S. (2015). Fault detection using random forest similarity distance. *IFAC-PapersOnLine*, 48(21), 583–588.

Verikas, A., Gelzinis, A., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern recognition*, 44(2), 330–349.

Wold, H. (1983). Systems analysis by partial least squares.

Wu, D., Jennings, C., Terpenny, J., Gao, R.X., and Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering*, 139(7).

Zhang, J. and Yang, C. (2014). Evaluation model of color difference for dyed fabrics based on the support vector machine. *Textile Research Journal*, 84(20), 2184–2197.