L. Parrenin, C. Danjou, R. Beauchemin, B. Agard, Prédiction de la qualité de la farine après mouture, 14ème Conférence Internationale Génie Industriel QUALITA – CIGI 2021, Grenoble, France, 5-7 mai 2021.

Prédiction de la qualité de la farine après mouture

LOIC PARRENIN [1,2], CHRISTOPHE DANJOU [1,2], ROBERT BEAUCHEMIN [3], BRUNO AGARD [1,2]

[1] Laboratoire en Intelligence des Données
Département de mathématiques et génie industriel,
École Polytechnique de Montréal, CP 6079, succursale Centre-Ville, Montréal, Québec, Canada
loic.parrenin@polymtl.ca, christophe.danjou@polymtl.ca, bruno.agard@polymtl.ca

[2]Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)
[3] La Meunerie Milanaise, Saint-Jean-sur-Richelieu, Québec, Canada

***Résumé* – Dans la production de farine biologique, différentes qualités de farine sont exigées par les industriels ou artisans en fonction du type de produit à fabriquer. La qualité de ces grains, matière première vivante, varie au cours du temps en fonction de nombreux facteurs, entrainant par conséquent des variations sur la qualité de la farine biologique. Pour obtenir la qualité de farine souhaitée, il est possible de réaliser sur la chaine de transformation des mélanges de blé ainsi que des ajustements opérés sur les machines de productions. L'article a pour objectif de proposer un outil de prédiction de la qualité de la farine produite, à travers un modèle de classification. Le modèle utilise les données collectées sur les mélanges ainsi que la qualité des grains de blé disponibles en amont de la chaine de production. L'outil permet d'identifier les caractéristiques influentes sur la qualité de la farine et d'anticiper certains choix dans la transformation des grains de blé.**

Abstract - In the production of organic flour, different qualities of flour are required by industrialists or craftsmen depending on the type of product to be produced. The quality of these grains, a living raw material, varies over time depending on many factors, thus leading to variations in the quality of organic flour. In order to obtain the desired flour quality, it is possible to make wheat mixes on the processing line as well as adjustments on the production machines. The article aims to propose a tool for predicting the quality of the flour produced, through a classification model. The model uses the data collected on the mixes as well as the quality of the wheat grains available upstream of the production chain. The tool allows to identify the characteristics that influence flour quality and to anticipate certain choices in wheat grains processing.

***Mots clés* – Industrie 4.0, Grains biologiques, Prédiction, Qualité, Classification.**
***Keywords* – Industry 4.0, Organic grains, Prediction, Quality, Classification**

1 INTRODUCTION

L'accroissement de la population mondiale et l'augmentation des niveaux de vie entrainent des besoins alimentaires plus importants. L'agriculture biologique est une méthode spécifique de production alimentaire utilisant des substances et des procédés naturels. Cette méthode encourage une utilisation responsable des ressources naturelles, le maintien et l'amélioration de la biodiversité, la qualité de l'eau et la fertilité des sols (MAPAQ, 2019). Le secteur céréalier biologique affiche en 2017 une croissance significative de la quantité de surface dédiée à la culture de céréales biologiques dans le monde avec un gain de 6% (représentant 280 000 hectares) par rapport à 2016 (Willer & Lernoud, 2019). D'après (Future Market Insights, 2020), on note une augmentation de la demande de farine de blé biologique, en Amérique du Nord et dans les pays d'Europe occidentale, notamment chez les fabricants de produits alimentaires, en raison de consommateurs plus soucieux de leur santé.

La farine est produite à partir du broyage de grains de blé. Le broyage (ou mouture) se déroule typiquement dans un moulin. L'objectif est de séparer les différentes parties du grain de blé et de les réduire en plus petits morceaux. Deux principales techniques existent pour la mouture du blé (Cappelli, Oliva & Cini, 2020) : la mouture sur meule et la mouture sur cylindre. Chaque technique présente ses avantages et inconvénients. Toutefois, la mouture sur cylindre est une technique plus récente et la plus répandue actuellement dans les moulins pour des raisons de rendements élevés. Elle permet d'obtenir des farines très blanches.

Il existe différentes catégories de blé :

- Le blé dur et le blé tendre sont des blés de dureté différente. Le blé dur possède une plus grande concentration en gluten, ce qui le rend très dur. Sa flexibilité et sa force en font un blé idéal pour les pâtes alimentaires. Le blé tendre est le blé le plus semé et le plus utilisé dans l'alimentation. Il est cultivé pour faire de la farine panifiable.
- Le blé de printemps et le blé d'hiver sont des blés semés à différentes périodes de l'année en fonction des zones climatiques. Un blé semé au printemps sera dit « blé de printemps » tandis qu'un blé semé à l'automne sera dit « blé d'hiver ».

Chaque type et variété de blé présente des caractéristiques différentes. De plus, les propriétés intrinsèques d'un type de grains de blé varient énormément entre les régions, les conditions météorologiques et climatiques, les périodes de récoltes et la rotation des cultures (Borghi et al., 1995; Johansson & Svensson, 1998; Triboi et al., 2000).

La production industrielle de farine biologique est quelque peu différente de la farine conventionnelle. Pour la production de farine conventionnelle, les volumes de production sont très importants, et très peu de mélanges sont réalisés. De plus, pour maintenir une qualité de production stable tout au long de l'année, il est possible d'avoir recours à des agents chimiques tels que le chlore ou le peroxyde de benzoyle pour le blanchissement de la farine, l'acide ascorbique pour améliorer la qualité boulangère et la couleur de la farine, ainsi que des enzymes pour uniformiser l'indice de chute (Gouvernement du Canada, 2020). Alors que pour la production de farine de blé biologique, les grains de blé utilisés sont cultivés sans aucune utilisation de produit chimique, de pesticides et d'engrais synthétiques (MAPAQ, 2019). Lors de la transformation, aucun ajout d'agent chimique n'est autorisé au regard de la production alimentaire biologique. Il faut alors utiliser la bonne proportion de différentes sortes de grains de blé, qui ensemble permettront d'obtenir le produit final voulu.

Dans ce contexte, **comment produire de la farine de blé biologique, de qualité constante tout au long de l'année, lorsque la qualité des matières premières fluctue selon la période de l'année, en fonction des conditions climatiques et météorologiques ?**

Dans le présent article, nous proposons une méthodologie qui vise à prédire différentes caractéristiques de la farine après mouture, en fonction de la qualité des grains en entrée et des proportions des grains de blé pour chaque farine produite. La méthodologie proposée s'appuie sur un arbre de décision.

La structure de l'article se décompose de la manière suivante. Dans l'état de l'art (section 2), nous présentons les éléments essentiels de la production de la farine (2.1), notamment les caractéristiques intrinsèques des grains de blé (2.1.1), les principales caractéristiques recherchées de la farine (2.1.2) et les principales étapes de la production de farine (2.1.3). Puis, l'état de l'art présente les outils existants dans le domaine de la meunerie pour prédire la qualité de la farine (2.2). Finalement, nous présentons ensuite les éléments essentiels de l'outil de prédiction utilisés (2.3). La section 3 présentera l'outil proposé. La section 4 montrera une validation de l'outil dans un contexte réel. Finalement, la section 5 rappelle les résultats et limites de la méthode proposée et propose quelques perspectives.

2 ÉTAT DE L'ART

Bien que le domaine de la transformation des céréales en général soit très bien documenté, il n'existe que très peu de références scientifiques qui décrivent la production de la farine biologique.

2.1 Production de la farine biologique

D'après une étude sur la production alimentaire biologique, la plus grande superficie consacrée aux produits céréaliers dans le monde était en 2017 pour le grain de blé (32%) suivi des grains de maïs (14%) (Willer & Lernoud, 2019). La demande de farine biologique augmente, ceci est d'autant plus vrai en période de crise liée au covid-19 (Fournier, 2020).

2.1.1 Caractéristiques intrinsèques des grains de blé

Différentes variétés de grain de blé existent et sont définies par des caractéristiques physiques spécifiques. Tel que schématisé dans la figure 1, le grain de blé est situé dans un épi de blé et se compose de 4 parties distinctes : le **germe** (≈ 3 % de la masse totale du grain); l'albumen ou **amande farineuse** ou endosperme (84%) ; les **enveloppes** (13%) et la brosse (masse négligeable).

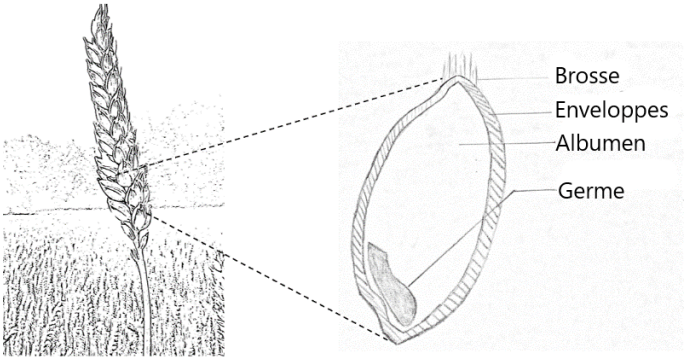

Figure 1. Schéma de l'anatomie d'un grain de blé (coupe longitudinale d'un grain de blé)

Le germe représente la future plante, elle est riche en matière grasse et en vitamines B et E. L'amande farineuse est principalement constituée de glucides (amidon) et de protéines (dont le gluten). L'amidon est un glucide complexe (sucre lent) utile à la croissance de la nouvelle plante. Le **gluten est une protéine** qui **donne aux pains l'élasticité** suffisante pour assurer un gonflement du pain. Les caractéristiques de chaque variété de blé varient en fonction du lieu de récolte et de la saison.

Pour évaluer la qualité d'un grain de blé, des analyses sont réalisées sur des échantillons. Les analyses mesurent notamment le **taux de protéine**, le **taux d'humidité**, **l'indice de chute**, le **taux de cendres**, le **BEM** (Brabender Energy Max) et le **PMT** (Peek Max Time). Le taux de protéine indique le pourcentage de protéine dans le grain. Le taux d'humidité précise la quantité d'eau présente dans le grain. L'indice de chute mesure l'activité enzymatique du grain qui se développe dans le grain dès le début de la germination (Côté, 2018), cette mesure informe sur le stade de germination du grain. Le taux de cendre mesure la quantité de minéraux présents dans un grain ou dans la farine. Le BEM et PMT sont des mesures obtenues par le test Glutopeak. L'appareil analyse les capacités d'agrégation du gluten en phase liquide (Gresle, 2013). Le BEM définit le couple maximum exprimé en unité brabender (UB) (Gresle, 2013) et précise la force boulangère de la pâte. Le pic de la courbe obtenue par le Glutopeak correspond au BEM (Gresle, 2013). Le PMT indique la durée nécessaire pour atteindre le couple maximum (Gresle, 2013).

2.1.2 Caractéristiques recherchées de la farine

Plusieurs types de farine existent en fonction du produit alimentaire final recherché : pâtisserie, biscuit, pain, pâte à pizza, pâte alimentaire, semoule…

Un premier choix est réalisé au niveau de la **dureté du blé**. Un blé dur est favorable pour produire une farine dédiée à la production de semoule ou de pâte alimentaire. Tandis qu'un blé tendre est utilisé pour produire de la farine panifiable (farine dédiée à la production de pain, pâtisserie, biscuit).

Au niveau du blé tendre, un premier paramètre de qualité s'oriente autour de **la proportion en gluten** (une protéine) présent dans le grain de blé. Un blé tendre plus riche en protéine sera adapté pour la fabrication de pain. Une farine plus riche en protéine, et donc en gluten, permettra d'obtenir les qualités d'élasticité et

d'extensibilité de la pâte lors de la confection du pain. Tandis qu'un blé tendre plus faible en protéine sera plus approprié pour la pâtisserie et les biscuits, où les propriétés d'une pâte cassante et d'absorption d'eau plus faible sont recherchées.

Un deuxième paramètre concerne la **proportion de « son »** contenu dans la farine. Cette proportion est mesurée à l'aide de l'indicateur du taux de cendre et constitue le principal critère de classement des farines de blé pour le grand public. Plus le taux de cendre est élevé, plus la farine contient du son. Le son regroupe l'enveloppe et le germe du blé. Une farine dite intégrale contient tous les éléments nutritifs présents dans un grain de blé (enveloppe, albumen, germe). Une farine complète contient seulement les parties de l'enveloppe et de l'albumen du grain de blé. Tandis qu'une farine blanche sera composée seulement de la partie de l'albumen.

Comme on a pu le mentionner, plus le taux de protéine est élevé, plus la proportion en gluten dans la farine est importante. Toutefois, le **taux de protéine ne reflète pas la qualité du gluten**. C'est pour cela que le BEM représente une mesure de la qualité du blé plus fiable que le taux de protéine. Le BEM semble être corrélé au **taux d'absorption** en eau qui est un paramètre important pour déterminer les propriétés fonctionnelles de la farine (Fu, Wang & Dupuis, 2017). Une farine avec une plus grande absorption en eau, est préféré pour la production du pain (propriété de manipulation de la pâte améliorée, pain de meilleure qualité au niveau du gout et de la structure) - (Fu et al., 2017). Tandis qu'avec un taux d'absorption plus faible, il est préférable de produire de la pâtisserie et des biscuits.

2.1.3 Les étapes de production de la farine de blé

La transformation des grains de blé en farine se déroule au moulin en plusieurs étapes. Les principales étapes de la production de farine de blé sont représentées sur la figure 2.

(1) **Les blés** achetés sont **livrés** par lot, par camions, au cours de l'année. Les blés développent des caractéristiques qui leur sont propres en fonction de leur environnement et de leur type. La grandeur des champs peut être également une source de variabilité de la qualité du blé avec des zones qui ont possiblement pu être moins irriguées que d'autres. Pour s'assurer de la qualité des blés réceptionnés, une prise d'échantillon des grains de blé est réalisée pour être ensuite analysée au laboratoire. Les résultats d'analyse permettent d'identifier le blé et de sélectionner le silo dans lequel les grains de blé seront entreposés. Afin de préserver la qualité de chaque type de blé et contrôler à un certain degré les mélanges, les blés de qualité similaire sont entreposés dans un même silo. La présence de multiples silos permet d'entreposer davantage de grain de blé, mais également de séparer les grains de blé de qualité différente. (2) Avant que les grains de blé soient entreposés, ils subissent une **étape de pré-nettoyage** (les criblures de blé désignent les impuretés) de manière à éliminer les gros déchets (paille, terre) ainsi que les objets métalliques pouvant endommager les machines de production.

Le talent du meunier consiste à bien **choisir les blés et les mélanger dans de bonnes proportions** afin de réaliser de bon mariage de blés et obtenir le type de farine souhaité, **ces mélanges correspondent à des recettes**. Les recettes sont réalisées en fonction de la qualité et de la disponibilité des matières premières. La recette définie, (3) les grains de blé sont **nettoyés**, l'objectif est de retirer toutes les impuretés (grains cassés, pierres, pailles…) pour ne garder que les grains de blé. (4) Les grains nettoyés passent au processus de mouillage dans le but d'**humidifier** les grains de blé et faciliter la séparation de l'albumen du reste des composants lors du processus de mouture. Les grains mouillés sont finalement envoyés au processus de mouture. L'opération de mouture **permet de séparer les différentes parties du grain de blé et de réduire en farine l'amande farineuse** du grain pour obtenir de la farine. La mouture se décompose en 4 étapes : le broyage ; le claquage ; le convertissage et le blutage. (5) **Le broyage** permet d'ouvrir le grain, à l'aide de cylindres cannelés, et d'obtenir de grosses particules des différentes parties du grain (semoules). (7) **Le claquage** réduit les semoules blanches (grosses particules d'amande farineuse). (8) **Le convertissage** réduit les semoules vêtues (particules dont une partie de l'amande farineuse est toujours collée à l'enveloppe). La réduction des particules se fait à l'aide de cylindres lisses. (6) À chaque étape, les particules sont triées en fonction de leur granulométrie, à partir de plusieurs tamis présents dans un *plansichter* pour être redirigées vers une nouvelle étape de broyage, de réduction ou dans un silo de produit transformé (farine ou son), c'est **l'étape de blutage**. Ces étapes sont réalisées de manières successives, d'où la présence dans le moulin de plusieurs machines de broyage, de claquage et de convertissage.

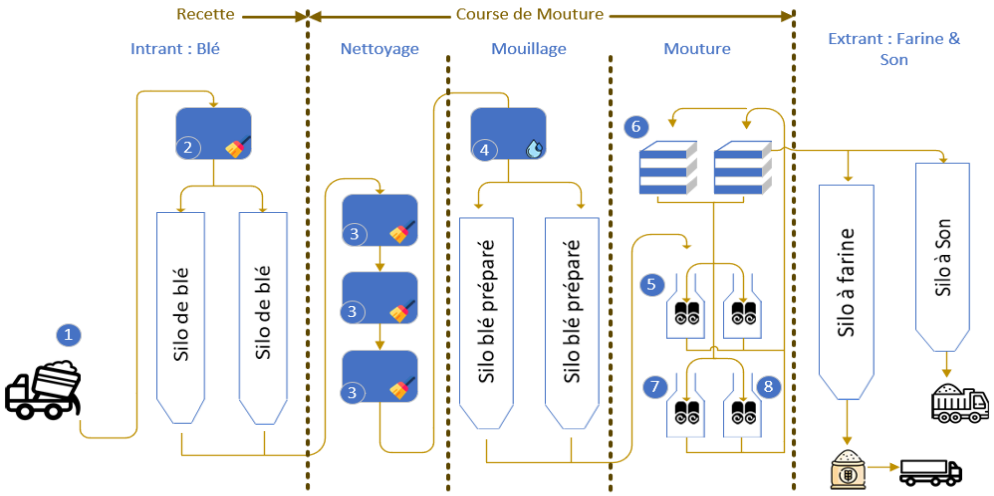

Figure 2. Processus de production de la farine de blé

Une multitude de paramètres doivent être contrôlés et certains ajustés en conséquence, tant au niveau des recettes, du mouillage que de la mouture afin de produire de la farine biologique. Cependant, l'influence et l'importance de chacun des paramètres sur l'ensemble du processus rendent leurs ajustements complexes.

2.2 Prédire la qualité de la farine

Actuellement dans le secteur de la farine, le meunier a pour rôle de sélectionner les mélanges de blés en entrée du processus de transformation et d'ajuster les paramètres de production pour obtenir une qualité de farine souhaitée. La qualité de la farine est mesurée généralement à l'aide d'un farinographe (temps de stabilité), alvéographe (index alveographique (W, P/L)) et d'un proche infrarouge (protéine, humidité, taux de cendre) - (Cocchi et al., 2005). Des moyennes sur la qualité des grains de blé mélangé ou un échantillon représentatif d'une recette peuvent être analysés avant de débuter une course de mouture de manière à estimer la qualité de la farine produite.

(Fu et al., 2017) montre que le couple max (BEM) du « Glutopeak » et le taux d'absorption d'eau de la farine présentent une forte corrélation positive avec un coefficient de détermination (r^2) de 0.97 sur 83 échantillons de grains de blé différents. Le test Glutopeak est donc être un outil rapide et utile pour prédire le taux d'absorption d'eau de la farine et déterminer sa qualité.

Des méthodes d'apprentissage machine sont aujourd'hui utilisées pour prédire la qualité ou les paramètres de certaines transformations d'aliment (Adebayo, Hashim, Abdan, Hanafi & Mollazade, 2016; Bao et al., 2014; Barbon et al., 2016; El-Bendary, El Hariri, Hassanien & Badr, 2015; Liu, Yang & Deng, 2015).

Dans le milieu de la farine, de nombreuses méthodes de prédiction existent également pour prédire la qualité de la pâte (Torbica, Blazek, Belovic & Hajnal, 2019) ou l'usage final de la farine (Cocchi et al., 2005) en fonction de sa qualité. Des prédictions de la qualité de la farine ont été réalisées à l'aide d'un modèle de réseaux neurones en utilisant comme donnée d'entrée les résultats (ondes) de la spectroscopie du proche infrarouge de différentes farines (Mutlu et al., 2011). L'objectif est différent du travail présenté ici, car (Mutlu et al., 2011) traitent des données d'analyse de farine pour prédire les caractéristiques de la farine (taux de protéine, taux d'absorption, etc.) dans le but de simplifier les analyses et réduire les temps d'analyse, alors que la présente étude traite des données d'analyses de grains de blé pour prédire une classe de qualité de farine.

Parmi les nombreuses méthodes étudiées, aucune ne fait le lien entre la qualité des matières premières et les recettes de transformation réalisées. La qualité des grains de blé varie au cours du temps (transport, entreposage) comme tout produit alimentaire interagissant avec son environnement extérieur. Toutefois des grains de blé de qualité différente sont parfois mélangés pour obtenir une certaine qualité de farine désirée. De plus, ces grains de blé subissent des opérations de nettoyage, mouillage et de mouture impactant la qualité de la farine. La qualité de la farine dépend donc de nombreux paramètres dans l'ensemble du processus de production. La manipulation de matière première vivante le long du processus de production rend la prédiction de la qualité de la farine complexe.

2.3 Outil de prédiction

Il existe différents outils de prédiction, dont les réseaux neurones (RN). Les RN sont devenus une technique populaire dans les sciences biologiques en raison de leur qualité prédictive (Alvarez, 2009; Dubey, Bhagwat, Shouche & Sainis, 2006). Les réseaux neurones ont pu être utilisés avec succès dans l'industrie alimentaire tels que : la classification de grains de céréales à l'aide de caractéristiques morphologiques (Visen, Paliwal, Jayas & White, 2002), la prédiction de la conductivité thermique des produits boulanger (Sablani, Baik & Marcotte, 2002), la prédiction de paramètres de qualité de la farine à partir des spectres du proche infrarouge (Mutlu et al., 2011), la prédiction de la qualité des petits pois en fonction du temps de cuisson (Xie & Xiong, 1999), etc.

Bien que les réseaux neurones offrent de bons résultats de prédiction, ce type de modèle a besoin d'avoir de nombreuses données d'entrainement se définissant par des données traitées et utilisées pour l'apprentissage du modèle, de manière à atteindre un niveau de prédiction précis (Wuest, Weimer, Irgens & Thoben, 2016). De plus, le modèle des réseaux neurones est difficilement interprétable, souvent caractérisé comme une « boite noire » qui n'explique pas les relations et l'influence des paramètres sur les résultats de prédiction (donnée de sortie). Un modèle de régression non linéaire ou d'**arbre de décision** semblent de bonnes alternatives au modèle des réseaux neurones afin de pouvoir **interpréter et comprendre ces modèles de prédiction et de classification**.

Un arbre de décision est un graphique orienté acyclique utilisé pour prendre des décisions (Burkov, 2019a). Il est composé de nœuds, de banches et de feuilles. Un nœud teste un attribut. Une branche correspond à la valeur d'un attribut suite à un test réalisé au niveau d'un nœud. Une feuille indique la classe à laquelle l'exemple appartient. L'arbre de décision permet à partir des données d'entrainement de tester plusieurs attributs et d'établir certaines règles, précisées à chaque nœud, pour classifier chaque exemple. Les données tests qui sont de nouvelles données indépendantes des données d'entrainement valideront le modèle de classification suivant les règles établies.

Dans une approche similaire, (Ronowicz, Thommes, Kleinebudde & Krysinski, 2015) ont étudié la qualité des pastilles pharmaceutique. Pour cela, ils ont exploré à l'aide d'un modèle d'arbre de décision, les relations de cause à effet entre les caractéristiques de la formulation (composition des granulés et des paramètres du procédé) et la sphéricité des granulés (attribut de qualité sélectionné du produit final). La découverte de ces relations de cause à effet entre les différentes variables offre la possibilité d'optimiser le processus de fabrication des granulées. Cette étude montre la pertinence d'une approche par arbre de décision dans le milieu pharmaceutique.

3 OUTIL DE PREDICTION DE LA FARINE A PARTIR DE LA QUALITE DES GRAINS DE BLE

L'État de l'art montre que les grains de blé sont une matière « vivante » définie par des propriétés intrinsèques évoluant avec le temps. La connaissance des caractéristiques de la farine produite est importante pour déterminer l'usage final du produit et répondre précisément au besoin du client.

Dans le milieu biologique, l'ajout d'agents chimiques pour atteindre les caractéristiques recherchées ne peut être envisageable au regard des normes (MAPAQ, 2019). Pour cela il est nécessaire de travailler avec la qualité des grains de blé récolté pour obtenir une qualité de farine donnée. Cependant, d'après l'état de l'art (section 2.2), il n'y a actuellement pas d'outils pour prédire la qualité de la farine à partir des matières premières. Le problème fait intervenir de nombreux facteurs impactant pour chacun d'entre eux la qualité de la farine à des degrés différents. La pratique se base actuellement sur l'expérience du meunier.

Nous proposons de **développer un outil de prédiction** afin de déterminer les caractéristiques de la farine qui sera produite après le processus de mouture. L'outil de prédiction utilise la moyenne de la qualité des grains de blé présent dans un silo (protéine, humidité, taux de cendre, BEM, PMT) ainsi que les pourcentages des lots de grains de blé de chaque silo. À partir de ces données d'entrées, l'outil indique la classe de la qualité de la farine associée au BEM de la farine.

Pour construire l'outil de prédiction, une méthode telle que représentée sur la figure 3 est suivie. Cette méthode se compose de 4 phases.

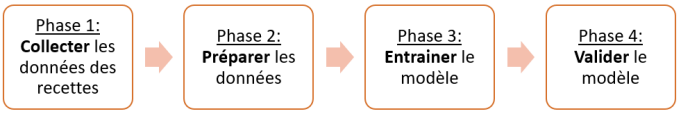

Figure 3. Méthodologie de prédiction

3.1 Phase 1 : collecter les données des recettes

Dans cette première phase, il est nécessaire de **définir les besoins**. Cette définition précise les données utiles pour le développement de l'outil de prédiction et délimite notre zone d'étude. Dans notre cas, nous souhaitons collecter les données précisant la qualité des grains présents dans chaque silo ainsi que leur quantité pour chaque recette réalisée.

La qualité des grains de blé est mesurée à l'aide d'un proche infrarouge (protéine, humidité, taux de cendre) et d'un test Glutopeak (BEM, PMT). Dans chaque silo, différents types de grain de qualité similaire peuvent être mélangés. Pour cela des moyennes sur le taux de protéine, le taux d'humidité, le taux de cendre ainsi que le BEM et le PMT sont calculés pour préparer une recette. Il est alors nécessaire de collecter les moyennes de la qualité des blés pour chaque silo ainsi que les pourcentages des quantités de blé de chaque silo utilisé pour la recette.

3.2 Phase 2 : préparer les données

Afin d'entrainer le modèle, il est nécessaire de préparer les données. Pour cela, nous rassemblons l'ensemble des données dans une **table de donnée** (« dataframe »). Un aperçu de l'entête de la table de donnée est montré sur le tableau 1.

RECETTE (R)	**A1**	**A2**	**…**	**Y**
R1	$a_{1,1}$	$a_{1,2}$	…	y1
R2	$a_{2,1}$	$a_{2,2}$	…	y2
R3	$a_{3,1}$	$a_{3,2}$	…	y3

Tableau 1. Table de donnée

Chaque ligne de cette table représente une **recette réalisée (R)**. Nous retrouvons ainsi en colonne, les informations collectées au niveau des recettes précisant la qualité des grains et les pourcentages des mélanges utilisés. Les attributs désignent les **caractéristiques mesurées (A)** pour identifier la qualité des grains de blé. À partir des proportions de quantité de blé utilisées pour chaque silo, une **moyenne sur chaque caractéristique ($a_{i,j}$)** est calculée. La **donnée de sortie (Y)** indique la classe de la qualité de la farine. La classe est identifiée à partir du BEM provenant des résultats d'analyses du test du Glutopeak. L'analyse est réalisée sur un échantillon de farine après le processus de mouture. La donnée de sortie doit être associée à la recette réalisée pour être en mesure d'entrainer le modèle. Ce type d'apprentissage est un **modèle d'apprentissage supervisé**, c'est-à-dire que les données sont annotées par une étiquette (« label ») afin que l'algorithme puisse être capable, une fois entrainé, de prédire l'étiquette sur de nouvelles données non annotées.

Pour certains modèles, il est nécessaire de **normaliser** les valeurs de manière à éviter qu'une entité domine sur les autres en raison d'une plus grande distribution ou plage de valeur (Burkov, 2020). Les arbres de décision ne sont pas sensibles face à ces disparités entre chaque entité, pour cela les données n'ont pas été normalisées.

Dans certains cas, il est possible que les **données** soient **débalancées** et que les valeurs de **certains attributs ne soient pas disponibles ou manquent dans l'ensemble de données** (Wuest et al., 2016). En fonction du modèle algorithmique choisi et du problème, différentes solutions existent et sont disponibles dans la littérature scientifique (Lakshminarayan, Harp, Goldman & Samad, 1996; Marlin, 2008; Zhang & Hu, 2014). Dans notre cas, en utilisant un modèle d'arbre de décision pour la prédiction et de la classification, le débalancement a un impact sur la structure de l'arbre décisionnel. Il est possible de rebalancer les données en infligeant une pénalité pour certaines erreurs de classification. La **classification pénalisée** impose un cout supplémentaire au modèle pour les erreurs de classification commises sur la classe minoritaire de l'ensemble de données (Datascientest, 2020). Ces pénalités biaisent le modèle afin qu'il accorde une plus grande importance à la classe minoritaire.

Parfois les ensembles de données présentent des données manquantes pour certains attributs, le plus souvent lorsque les données sont rentrées manuellement. Plusieurs approches sont possibles : supprimer les recettes dont les données sont manquantes ; utiliser un modèle d'apprentissage pour remplacer les valeurs manquantes ; utiliser une technique d'imputation des données. Avec un ensemble de données suffisamment grand, ne contenant peu de données manquantes, il a été choisi de **supprimer les recettes ayant des données manquantes** dans le but de garder des informations précises et fiables.

Pour être en mesure d'entrainer et valider notre modèle, nous séparons notre ensemble de données en **trois sous-ensembles**. On utilise **70%** des données pour les **données d'entrainement**, **15%** pour **la validation** et **15 %** pour **le test** (Burkov, 2019b). Ces proportions varient en fonction de la dimension des données (Burkov, 2019b). Afin d'entrainer correctement notre modèle et éviter de créer des biais sur notre ensemble de données tests, nous gardons une proportion représentative de chaque classe de sortie dans les 3 sous-ensembles. Pour cela, il est possible de réaliser un échantillonnage **stratifié**.

3.3 Phase 3 : entrainer le modèle

Plusieurs algorithmes d'arbres de décision existent : ID3, C4.5, C5.0, CART (Classification and Regression Trees) - (scikit-learn developers, 2020). L'algorithme d'arbre de décision le plus connu et utilisé est l'algorithme C4.5. Il accepte les attributs discrets et continus, gère les exemples incomplets et résout les problèmes de surapprentissage par « élagage ». **L'outil de prédiction utilise une version optimisée de l'algorithme CART à l'aide de la librairie scikit-learn** sous le langage Python (Pedregosa et al., 2011).

Lors de l'entrainement du modèle, il est important d'éviter d'avoir un surapprentissage ou sous-apprentissage du modèle. On parle de surapprentissage si la performance du modèle sur les données d'entrainement est considérablement meilleure que sur les données de test. Lorsque le modèle fait du **surapprentissage, la variance est élevée** car le modèle aura appris le bruit présent dans les données d'entrainement. Le bruit sont des données ayant des fluctuations aléatoires ou des décalages par rapport aux valeurs réelles au niveau des attributs et de la variable de réponse. Ce bruit peut masquer la relation réelle entre les attributs et la variable de réponse. Dans le cas où le modèle réalise de nombreuses mauvaises prédictions sur les données d'entrainement, le modèle fait du **sous-apprentissage et possède un biais élevé**. Les raisons peuvent être multiples : un modèle trop simple pour les données étudiées ; un nombre d'attributs qui ne fournit pas suffisamment de données assez informatives pour enrichir le modèle et représenter la réalité (Burkov, 2019b).

Pour optimiser la performance du modèle, il est nécessaire d'affiner les paramètres de l'algorithme. Ces paramètres appelés hyperparamètres sont ajustés de manière expérimentale. Avec un modèle d'arbre de décision, il est possible d'ajuster les hyperparamètres par des critères d'arrêt (pré-élagage) ou des méthodes post-élagage. Les critères d'arrêt regroupent des techniques telles que : limiter la profondeur maximale de l'arbre; préciser le nombre minimum d'exemples nécessaires pour réaliser une séparation; indiquant le nombre minimum d'objets nécessaires pour avoir une feuille. Les méthodes post-élagage sont par exemple : fixer une valeur minimale à atteindre pour la diminution de l'impureté lors d'une séparation à un nœud; ajouter une fraction de poids à respecter au minimum pour la création d'une feuille et de ses branches**. Le bon ajustement de ces paramètres doit permettre de trouver un juste équilibre entre le biais et la variance.**

De plus, face à des données débalancées au niveau de la variable de sortie, une classification pénalisée est réalisée comme décrit dans la phase 2. Une analyse de sensibilité du cout en fonction de la précision (« accuracy ») est examinée afin de déterminer les pénalités optimales à attribuer pour chaque type de mauvaise classification.

3.4 Phase 4 : valider le modèle

À partir de l'ensemble de **données tests**, il sera possible de valider la performance du modèle créé et ajusté avec les données d'entrainement et de validation. L'évaluation du modèle de classification est réalisée à partir d'une **matrice de confusion** de manière à connaitre les résultats de classification et la précision du modèle à prédire la bonne classe de sortie. La matrice de confusion renseigne sur les bonnes classifications, les faux positifs et les faux négatifs.

Le tableau ci-dessous montre une matrice de confusion 2x2, représenté par seulement 2 classes de sorties.

	Classe prédite		
Classe réelle		Classe 1	Classe 2
	Classe 1	*Vrai Positif*	*Faux Négatif*
	Classe 2	*Faux Positif*	*Vrai Négatif*

Tableau 2: Matrice de confusion 2x2

On mesure la performance du modèle à partir de la métrique de précision. La précision (« accuracy ») est calculée de cette manière :

$$Précision = \frac{\text{Vrai Positif} + \text{Vrai Négatif}}{\text{Vrai Positif} + \text{Faux Positif} + \text{Faux Négatif} + \text{Vrai Négatif}}$$

L'objectif de l'étude est de développer un outil de prédiction pour prédire la qualité de farine produite à partir de la qualité des grains de blé utilisée. La revue de littérature ne fournit pas d'outil semblable qui réponde à cet objectif. Pour cette raison, un outil utilisant des algorithmes d'apprentissages machines a été développé. L'outil de prédiction se base sur les données d'analyses d'échantillons de blé ainsi que des recettes pour classifier la qualité de la farine dans une catégorie adaptée pour certaines industries.

4 CAS D'ETUDE/VALIDATION

Afin de valider l'outil proposé, nous allons le tester avec un cas d'étude. Il n'existe présentement d'après la revue de littérature aucune méthode ou outil offrant des résultats en lien avec l'objectif étudié pour comparer les résultats de l'outil proposé. Pour cela, **la moyenne de la valeur du BEM des grains de blé pour chaque recette servira de référence**.

4.1 Contexte

L'outil a été testé sur les données d'une entreprise de transformation de grains céréaliers biologiques au Canada. L'entreprise connait depuis quelques années une forte croissance des ventes. Cette croissance à deux chiffres implique une augmentation de la production et passe par un volume de production plus important et/ou par une meilleure productivité. En raison de machine très couteuse et en cherchant à minimiser les pertes lors de la transformation ainsi que les temps de production, une meilleure productivité est recherchée. La prédiction de la qualité de la farine permettra éventuellement de **comprendre certains facteurs influençant la qualité finale de la farine produite** et **anticiper certains choix dans la transformation comme l'ajustement des recettes** pour atteindre la qualité souhaitée. Cette anticipation offrira des gains de productivité sur la chaine de production.

Les données mises à disposition par le partenaire industriel regroupent : les analyses d'échantillons de grains de blé réceptionnés à l'usine ; les recettes ; les analyses d'échantillons de farine prélevés à la fin de la mouture et les courses de mouture (provenant du système ERP – Entreprise Resource Planning).

Le partenaire industriel possède 12 silos pour l'entreposage de grain de blé. Chaque silo contient plusieurs tonnes de grains de blé pouvant être mélangées avec d'autres lots de grains de blé de fournisseurs différents. Les grains de blé de qualité similaire sont regroupés dans un même silo.

Chaque recette est sauvegardée dans un fichier Excel sous un dossier sur le serveur de la compagnie. Le fichier recette précise les valeurs des caractéristiques moyennes des grains de blé présent dans chaque silo, les quantités des grains de blé et proportions utilisées de chaque silo ainsi que la quantité totale de blé à utiliser. Les caractéristiques moyennes des grains de blé sont calculées automatiquement à partir du fichier Excel répertoriant les résultats d'analyses d'échantillons de grains de blé réceptionnés à l'usine.

4.2 Mise en œuvre de la démarche

4.2.1 Phase 1 : collecter les données des recettes

De nombreuses données sont enregistrées et disponibles à travers différentes interfaces et systèmes informatiques. Dans cette première phase, nos besoins se porteront sur les recettes ainsi que les données d'analyses réalisées au laboratoire pour identifier la qualité des blés ainsi que la qualité de la farine produite. La collecte de ces données et l'étiquetage de ces données sont primordiales pour avoir des exemples et fournir un contexte afin que le modèle d'apprentissage machine puisse en tirer des enseignements. L'étiquetage est rendu possible à partir du numéro de course. Chaque nouvelle recette est liée à un numéro de course créée automatiquement à l'aide du système ERP. Toutefois, certaines courses de mouture ne possèdent pas de recette si aucun mélange n'est nécessaire. Les liens entre les recettes et les numéros de courses n'étant pas sauvegardés sur le système ERP sont retrouvés à travers les courriels du partenaire industriel ainsi que des fichiers Excel.

L'ensemble des données au niveau des recettes, des analyses et des courses de production sont inscrites manuellement pouvant accroitre le risque de données manquantes ou erronées. Pour cela, après l'étiquetage des données, il est important de vérifier la cohérence des données avec plusieurs sources d'information en prélevant un échantillon de l'ensemble de données. En raison d'un temps d'étiquetage des données important, une durée de 6 mois est sélectionnée. **170 recettes sont collectées sur une période de 6 mois (janvier 2020 à juin 2020).**

4.2.2 Phase 2 : préparer les données

Dans la phase 2, nous rassemblons les données des recettes dans une table de données à l'aide d'un script Python. Le script lit chaque fichier Excel utilisant un même template pour collecter l'information présente au niveau des recettes et l'inscrire dans une table de données sous une ligne. La figure 4 montre une partie de l'entête de la table obtenue :

Figure 4. Table de donnée des recettes

La figure 4 présente la dimensionnalité de la table regroupant les 6 caractéristiques (A) des grains de blé ainsi que les proportions utilisées pour les 12 silos. Le numéro de course de mouture étant un numéro unique pour chaque recette est identifié comme index.

Le calcul de la moyenne de chaque caractéristique sur les 12 silos en fonction des proportions utilisées permet de réduire le nombre de colonnes de 87 à 9. La moyenne de chaque caractéristique ($a_{i,j}$) est calculée par cette formule :

$$\sum_{k=1}^{12} (\text{Silo k } A_j) * (Silo\ k\ Mélange) = a_{i,j}\ , j = 1, \dots, 6 \qquad (1)$$

Les recettes sont identifiées par la date et le type de farine. Ces deux données permettent de vérifier sur un échantillon d'exemple collecté l'exactitude des informations et d'ajouter le numéro de course de mouture associé à la recette. Le numéro de course permet ensuite l'étiquetage de la donnée de sortie sur l'ensemble de données. Finalement la date et le type de farine sont supprimés de la table de donnée et la donnée de sortie y est ajoutée afin que le modèle s'entraine uniquement à partir des données décrivant la qualité des grains de blé.

Dans l'ensemble de données, s'il n'est pas possible d'identifier la recette à son numéro de course en raison d'un manque d'information ou d'information non cohérente, les recettes sont supprimées de la table de données. Les doublons sont également retirés. Au total 15 recettes ont été supprimées de la table de donnée.

À partir de l'expertise du partenaire industriel, la qualité de la farine est répartie en 7 classes. Seulement en raison d'une forte disparité du nombre d'exemples présente dans chacune des classes, le nombre de classes a été réduit à 4. Les 4 classes sont notées sur une échelle de 0 à 3 avec la classe 0 représentant une farine possédant un BEM faible et la classe 3 représentant une farine possédant un BEM très élevé.

Finalement en présence d'un débalancement de nos données, 30 recettes identifiées spécifiquement pour les catégories débalancées (classe 0 et 3) sont ajoutées à l'ensemble de données. Ces nouvelles recettes ont été réalisées durant la période entre juillet et octobre 2020. Suite à la réduction des classes et de l'ajout des recettes, la répartition des exemples se présente comme suit : 32 pour la classe 0, 49 pour la classe 1, 63 pour la classe 2 et 37 pour la classe 3.

4.2.3 Phase 3 : entrainer le modèle

L'entrainement du modèle de l'arbre de décision est réalisé à l'aide de la librairie Scikit-learn sur python. L'ensemble de données préparé est divisé en deux ensembles de la manière suivante : **85% pour les données d'entrainement et 15% pour les données tests**. Cette division est réalisée par un échantillonnage stratifié de la variable de sortie. Les paramètres énoncés dans la section 3.3 ont été optimisés par validation croisée avec 5 groupes différents échantillonnés de l'ensemble de données d'entrainement. La **fonction « GridSearchCV » est utilisée pour tester une multitude de paramètres souhaités** rapidement et connaitre celles qui offrent la meilleure performance. Pour rebalancer l'ensemble des données, des poids sont ajoutés pour pénaliser la classification erronée sur des classes minoritaires (classe 0 et 3). Les paramètres optimaux obtenus pour le modèle sont : « max_depth » = 5 ; « min_samples_leaf » = 6 ; « min_samples_split » = 2 et « class_weight » = {'a' : 5}.

4.2.4 Phase 4 : valider le modèle

Le modèle d'apprentissage est validé à partir des données tests. La précision du modèle sur les données d'entrainement est d'environ 58% avec un écart type d'environ 5.6% entre les 5 groupes de la méthode de la validation croisée. Ce faible écart type indique que le modèle est assez robuste. **La précision du modèle sur les données de test est d'environ 64%**, ce qui est meilleur que la précision obtenue sur les données d'entrainement. Cela indique que le modèle ne fait pas de surapprentissage.

4.3 Analyse des résultats

Nous obtenons **un arbre à 5 niveaux avec 15 feuilles**. **Les 2 variables les plus importantes** d'après l'arbre obtenu sont **la moyenne du BEM** ainsi que **le taux de protéine moyen** de chaque recette. La figure 5 montre les résultats de classification du modèle à partir des données tests dans une matrice de confusion.

	precision	recall	f1-score	support
0	0.83	1.00	0.91	5
1	0.50	0.57	0.53	7
2	0.64	0.70	0.67	10
3	0.67	0.33	0.44	6
accuracy			0.64	28
macro avg	0.66	0.65	0.64	28
weighted avg	0.64	0.64	0.63	28

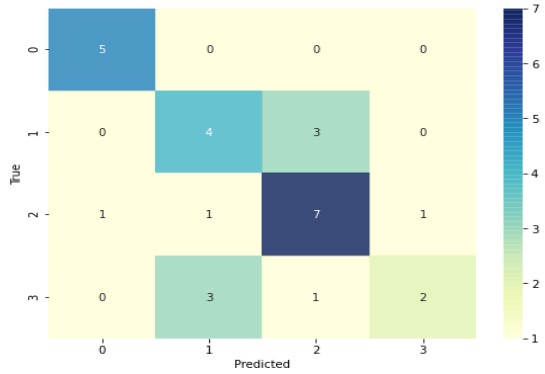

Figure 5. Matrice de confusion du modèle de l'arbre de décision

On observe d'après la figure 5 que le modèle classe très bien les farines de la classe 0 avec un BEM faible. Il classe relativement bien les farines de la classe 1 et 2. Toutefois les farines ayant un BEM élevé (classe = 3) sont assez mal classées et réalisent de mauvaises prédictions. Sur les 6 exemples présents appartenant à la classe 3, seulement 2 sont correctement prédits par le modèle.

En utilisant à partir des données tests seulement la moyenne des BEM (moyenne pouvant être calculée lors des recettes), la méthode offre une précision d'environ 57%, ce qui est légèrement inférieur aux résultats obtenus par l'arbre de décision.

Sur le cas d'étude, **l'outil de prédiction a permis de comprendre les variables d'intérêts telles que : le BEM, le taux de protéine, le taux d'humidité et le taux de cendre**. L'arbre permet également de visualiser et comprendre comment la classification est réalisée. Dans le cas d'étude, **le modèle performe légèrement mieux qu'une classification réalisée par la moyenne des BEM.** Pour de meilleures performances, d'autres modèles d'apprentissages tels que les réseaux neurones ou les arbres de type random forest peuvent être envisagés. Cependant, ces modèles se traduisent par une perte d'interprétabilité des résultats.

5 Conclusion

La qualité de la farine biologique varie tout au long de l'année selon la qualité des matières premières dont les caractéristiques changent au cours du temps en fonction des conditions environnementales. Pour cela, l'objectif de cet article était de proposer un outil capable de prédire la qualité de la farine après la mouture à partir des données de la qualité des grains présents dans des silos. Cet outil s'appuie sur 4 phases. Celles-ci sont (1) collecter les données, (2) préparer les données collectées, (3) entrainer le modèle d'apprentissage à partir des données préparées et finalement (4) valider le modèle. La démarche a été testée sur un cas d'étude. Les résultats ont été significatifs, car on note une performance légèrement supérieure à un simple calcul de la moyenne du BEM. Toutefois, on observe d'après les résultats la présence de mauvaises classifications dans chaque classe et plus particulièrement dans la dernière classe. Ces mauvaises classifications peuvent s'expliquer par un manque d'information au niveau des paramètres de production concernant la transformation des grains de blé. Le modèle pourrait être amélioré en incluant certains paramètres clés de production impactant la qualité de la farine (l'écartement des rouleaux d'écrasement, la durée de trempage des grains, le débit massique des grains acheminés vers les rouleaux d'écrasement sont des exemples de paramètres intéressants à intégrer au modèle).

En dehors de l'ajout de paramètres, une comparaison de la performance du modèle avec une approche neuronale entrainé sur un volume de donnée plus important semble pertinente. De plus, l'utilisation de l'ensemble des données de chaque silo, autrement que par le calcul de la moyenne pondérée de chaque caractéristique, doit être étudiée. La moyenne pondérée engendre une perte d'information, pouvant possiblement être utile à l'apprentissage du modèle et le rendre plus robuste.

La prédiction de la qualité de la farine offre la possibilité d'optimiser les recettes en termes de cout et de volume selon la qualité et quantité de grains de blé disponibles. Une piste qui sera étudiée plus en détail avec le partenaire industriel La Milanaise. La qualité de la farine est un paramètre important par la suite pour obtenir une fermentation optimale et des propriétés d'élasticité, de ténacité et d'extensibilité de la pâte propre à chaque client en fonction de la qualité du pain recherchée.

6 Remerciements

Nous tenons à remercier notre partenaire industriel, La Milanaise, d'avoir participé au développement de cet outil ainsi que le MAPAQ (projet IA119053) pour leur soutien financier.

7 References

Adebayo, S. E., Hashim, N., Abdan, K., Hanafi, M., & Mollazade, K. (2016). Prediction of quality attributes and ripeness classification of bananas using optical properties. *Scientia Horticulturae, 212*, pp. 171-182.

Alvarez, R. (2009). Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy, 30*(2), pp. 70-77.

Bao, Y. D., Liu, F., Kong, W. W., Sun, D. W., He, Y., & Qiu, Z. J. (2014). Measurement of Soluble Solid Contents and pH of White Vinegars Using VIS/NIR Spectroscopy and Least

Squares Support Vector Machine. *Food and Bioprocess Technology, 7*(1), pp. 54-61.

Barbon, A. P. A. C., Barbon, S., Mantovani, R. G., Fuzyi, E. M., Peres, L. M., & Bridi, A. M. (2016). Storage time prediction of pork by Computational Intelligence. *Computers and Electronics in Agriculture, 127*, pp. 368-375.

Borghi, B., Giordani, G., Corbellini, M., Vaccino, P., Guermandi, M., & Toderi, G. (1995). Influence of Crop-Rotation, Manure and Fertilizers on Bread-Making Quality of Wheat (Triticum-Aestivum L). *European Journal of Agronomy, 4*(1), pp. 37-45.

Burkov, A. (2019a). Chapiter 3: Fundalental Algorithms. The Hundred-Page Machine Learning Book. pp. 36-55

Burkov, A. (2019b). Chapiter 5: Basic Practice. The HundredPage Machine Learning Book. pp. 68-87

Burkov, A. (2020). Chapiter 4: Feature Engineering. Machine Learning Engineering. True Positive Inc. pp. 79-124

Cappelli, A., Oliva, N., & Cini, E. (2020). Stone milling versus roller milling: A systematic review of the effects on wheat flour quality, dough rheology, and bread characteristics. *Trends in Food Science & Technology, 97*, pp. 147-155.

Cocchi, M., Corbellini, M., Foca, G., Lucisano, M., Pagani, M. A., Tassi, L., & Ulrici, A. (2005). Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Analytica Chimica Acta, 544*(1-2), pp. 100-107.

Côté, M.-H. (2018). Évaluation de différents taux d'humidité de la récolte du blé panifiable tout en considérant la qualité boulangère et l'aspect économique. Groupe multiconseil agricole Saguenay-Lac-Saint-Jean.

Datascientest. (2020). *Classification pénalisée*. Tiré de: https://datascientest.com/glossary/classification-penalisee

Dubey, B. P., Bhagwat, S. G., Shouche, S. P., & Sainis, J. K. (2006). Potential of artificial neural networks in varietal identification using morphometry of wheat grains. *Biosystems Engineering, 95*(1), pp. 61-67.

El-Bendary, N., El Hariri, E., Hassanien, A. E., & Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Systems with Applications, 42*(4), pp. 1892-1905.

Fournier, M.-E. (2020). Vendre de la farine, ce n'est pas de la tarte! . *La Presse*. Tiré de: https://www.lapresse.ca/affaires/entreprises/2020-04-18/vendre-de-la-farine-ce-n-est-pas-de-la-tarte

Fu, B. X., Wang, K., & Dupuis, B. (2017). Predicting water absorption of wheat flour using high shear-based GlutoPeak test. *Journal of Cereal Science, 76*, pp. 116-121.

Future Market Insights. (2020). *Organic Wheat Flour Market: Global Industry Analysis 2012 - 2016 and Opportunity Assessment; 2017 - 2027*.

Règlement sur les aliments et drogues (C.R.C., ch. 870), (2020). Tiré de: https://laws-lois.justice.gc.ca/fra/reglements/c.r.c.,_ch._870/page-65.html

Gresle, E. (2013). Du Farinograph au Glutopeak. *Tribune des Constructeurs,* pp. 27-33. https://www.aemic.com/uploads/pdfs/IdC184.27.pdf

Johansson, E., & Svensson, G. (1998). Variation in bread-making quality: Effects of weather parameters on protein concentration and quality in some Swedish wheat cultivars grown during the period 1975-1996. *Journal of the Science of Food and Agriculture, 78*(1), pp. 109-118.

Lakshminarayan, K., Harp, S. A., Goldman, R., & Samad, T. (1996). *Imputation of missing data using machine learning techniques*. AAAI.

Liu, C., Yang, S. X., & Deng, L. (2015). A comparative study for least angle regression on NIR spectra analysis to determine internal qualities of navel oranges. *Expert Systems with Applications, 42*(22), pp. 8497-8503.

MAPAQ. (2019). *Agriculture biologique.* Tiré de: https://www.mapaq.gouv.qc.ca/fr/Productions/Production/agriculturebiologique/Pages/alimentsbio.aspx

Marlin, B. M. (2008). *Missing Data Problems in Machine Learning*, University of Toronto. Tiré de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.8408&rep=rep1&type=pdf

Mutlu, A. C., Boyaci, I. H., Genis, H. E., Ozturk, R., Basaran-Akgul, N., Sanal, T., & Evlice, A. K. (2011). Prediction of wheat quality parameters using near-infrared spectroscopy and artificial neural networks. *European Food Research and Technology, 233*(2), pp. 267-274.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, pp. 2825-2830.

Ronowicz, J., Thommes, M., Kleinebudde, P., & Krysinski, J. (2015). A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. *European Journal of Pharmaceutical Sciences, 73*, pp. 44-48.

Sablani, S. S., Baik, O. D., & Marcotte, M. (2002). Neural networks for predicting thermal conductivity of bakery products. *Journal of Food Engineering, 52*(3), pp. 299-304.

Scikit-learn developers. (2020). *Decision Trees*. Tiré de: https://scikit-learn.org/stable/modules/tree.html

Torbica, A., Blazek, K. M., Belovic, M., & Hajnal, E. J. (2019). Quality prediction of bread made from composite flours using different parameters of empirical rheology. *Journal of Cereal Science, pp. 89*

Triboi, E., Abad, A., Michelena, A., Lloveras, J., Ollier, J. L., & Daniel, C. (2000). Environmental effects on the quality of two wheat genotypes: 1. quantitative and qualitative variation of storage proteins. *European Journal of Agronomy, 13*(1), pp. 47-64.

Visen, N. S., Paliwal, J., Jayas, D. S., & White, N. D. G. (2002). AE—Automation and Emerging Technologies: Specialist Neural Networks for Cereal Grain Classification. *Biosystems Engineering, 82*(2), pp. 151-159.

Willer, H., & Lernoud, J. (2019). *The world of organic agriculture - Statitics & Emerging trends 2019*. IFOAM – Organics International & Research Institute of Organic Agriculture FiBL.

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research, 4*(1), pp. 23-45.

Xie, G., & Xiong, R. (1999). Use of hyperbolic and neural network models in modelling quality changes of dry peas in long time cooking. *Journal of Food Engineering, 41*(3-4), pp. 151-162.

Zhang, X. W., & Hu, B. G. (2014). A New Strategy of Cost-Free Learning in the Class Imbalance Problem. *Ieee Transactions on Knowledge and Data Engineering, 26*(12), pp. 2872-2885.