Lisbon School of Economics & Management
Universidade de Lisboa

MASTERS IN MANAGEMENT (MIM)

MASTERS FINAL WORK

PROJECT

MACHINE LEARNING APPROACHES FOR TARGETED DOUGH PROFILES IN INDUSTRIAL FLOUR PRODUCTION

DAVID MARIN BELTRAN

APRIL - 2026

MASTERS IN MANAGEMENT (MIM)

MASTERS FINAL WORK

PROJECT

MACHINE LEARNING APPROACHES FOR TARGETED DOUGH PROFILES IN INDUSTRIAL FLOUR PRODUCTION

DAVID MARIN BELTRAN

SUPERVISOR: PROF. CARLOS MANUEL JORGE DA COSTA

Jury:
PRESIDENT: PROF. JOSÉ MANUEL CRISTOVÃO VERISSIMO
RAPPORTEUR: PROF. FLÁVIO ALEXANDRE COSTA ROMÃO
SUPERVISOR: PROF. CARLOS MANUEL JORGE DA COSTA

APRIL - 2026

To Claudia, Oralia and
Natalia whose help and
patience always brought
the best in me.

GLOSSARY

Additives (Enzymes) - Substances added to flour to modify dough behaviour and improve processing performance. They represent key controllable decision variables in formulation.

Flour Characteristics - Intrinsic properties of flour (e.g., protein, moisture) that influence performance but are not directly controllable.

Flour Formulation - The process of defining additive types and dosages to achieve desired dough performance and meet operational requirements.

Mixolab Test - Industrial laboratory test used to evaluate dough behaviour under controlled mixing, heating, and cooling conditions.

Dough Profile - The measured behaviour of dough across the Mixolab test, used as a reference for quality and performance.

Stage-Dependent Modelling - Analytical approach that reflects the sequential nature of dough transformation by modelling each stage progressively.

Machine Learning Models - Data-driven tools used to predict outcomes and support decision-making based on historical data.

Inverse (Target-Oriented) Analysis - Approach that starts from a desired outcome and identifies input configurations needed to achieve it.

Decision-Support Tool - A system that enhances managerial and technical decision-making through data-driven insights.

Process Variability - Natural variation in materials and production conditions that affects consistency and requires management.

RESUMO

A produção industrial de farinha depende da utilização controlada de aditivos ou enzimas para garantir um desempenho consistente, apesar da variabilidade das matérias-primas. No entanto, as decisões de formulação são frequentemente baseadas em abordagens de tentativa e erro, o que limita a capacidade de tomada de decisão proativa. Esta investigação é orientada por uma questão central: em que medida modelos de machine learning dependentes de etapas podem apoiar a formulação e a tomada de decisão gerencial, reduzindo a incerteza na produção industrial de farinha.

Para responder a esta questão, o estudo segue três objetivos: (i) avaliar a capacidade preditiva de modelos de machine learning no comportamento da massa com base nas características da farinha e no uso de aditivos; (ii) identificar que etapas do processo reológico da massa são mais previsíveis; e (iii) avaliar se um enquadramento de modelização dependente de etapas pode apoiar a formulação direcionada ao informar a dosagem de aditivos antes dos testes laboratoriais.

A investigação adota uma abordagem orientada por dados com base na metodologia CRISP-DM, utilizando dados industriais de Mixolab provenientes de um produtor de farinha de grande escala. Modelos de Floresta de Decisão Aleatórias dependentes de etapas são desenvolvidos para refletir a natureza sequencial da transformação da massa e são avaliados através de métricas preditivas padrão. Para além da previsão direta, os modelos são aplicados num contexto inverso para explorar configurações de aditivos alinhadas com perfis de massa desejados.

Os resultados demonstram um forte desempenho preditivo na maioria das etapas, com menor precisão nas fases enzimáticas. De forma relevante, o enquadramento gera recomendações realistas de dosagem de aditivos, apoiando uma formulação mais proativa. Do ponto de vista gerencial, isto permite uma avaliação mais antecipada da viabilidade, reduz a incerteza e melhora a coordenação entre as áreas comercial, de I&D e de produção.

Palavras-chave: Machine Learning, Produção Industrial de Farinha, Modelagem Preditiva, Otimização de Aditivos, Controle de Processos

ABSTRACT

Industrial flour production depends on the controlled use of additives or enzymes to achieve consistent performance despite variability in resources. However, formulation decisions are often based on trial-and-error approaches which limit proactive decision-making. This research is guided by a central question: to what extent machine-learning–based, stage-dependent models can support formulation and managerial decision-making by reducing uncertainty in industrial flour production.

To address this question, the study pursues three objectives: (i) to assess the predictive capability of machine-learning models for dough behaviour based on flour characteristics and additive usage; (ii) to identify which stages of the dough rheology process are more reliably predictable; and (iii) to evaluate whether a stage-dependent modelling framework can support targeted formulation by informing additive dosing prior to laboratory testing.

The research adopts a data-driven approach based on the CRISP-DM methodology, using industrial Mixolab data from a large-scale flour producer. Stage-dependent Random Forest models are developed to reflect the sequential nature of dough transformation and are evaluated using standard predictive metrics. In addition to forward prediction, the models are applied in an inverse setting to explore additive configurations aligned with desired dough profiles.

The results show strong predictive performance for most stages, with lower accuracy in enzymatic phases. Importantly, the framework generates realistic additive recommendations, supporting proactive formulation. From a managerial perspective, this enables earlier feasibility assessment, reduces uncertainty, and improves coordination between commercial, R&D, and production functions.

Keywords: Machine Learning, Industrial Flour Production, Predictive Modelling, Additive Optimization, Process Control.

TABLE OF CONTENTS

FIGURES

TABLES

ACKNOWLEDGMENTS

First, I want to thank my family, thanks to my mother and grandmother I could reach every step on my academic and professional life.

Thank you, Professor Carlos, for your guidance throughout this project, without your help this would not have been possible.

Finally, I reiterate my thanks to my girlfriend who has been my support for the past two years while living abroad from our home country.

1. INTRODUCTION

In recent years, the production of flour and its transformation into finished goods such as bread, tortillas, or pasta have faced increasing scrutiny, not only from a technological point of view, but also from a sustainability and responsible production perspective (FAO, 2018). Worldwide initiatives such as the Sustainable Wheat Initiative, supported by the European bakery industry, show how reductions of waste are key concerns regarding the industry 2030 agenda (European Flour Millers & European Bakery Industry, 2020). Flour producers play a central role in the food supply chain as they oversee the milling, treatment, and enrichment of a raw material that ultimately reaches households and industrial producers. That is why ensuring that flour production is carried out responsibly is vital for food quality, safety, resource efficiency, and environmental consciousness (United Nations, 2015).

To highlight how widespread and significant flour production and consumption are globally, annual wheat utilisation was estimated at around 775 million tonnes in recent years (FAO, 2023). For example, major countries consume over 10 million tonnes of wheat per year, with China's consumption estimated at approximately 103 million tonnes, and India's just over 90 million tonnes (FAO, 2023). These figures illustrate that wheat flour and grain processing in general are not niche industries but rather massive global operations. Because of this scale, even moderate improvements in production efficiency, additive use, or waste reduction can lead to very large absolute impacts across the food system (United Nations, 2015).

A central dimension of flour production is the use of additives in the production process, among which enzymes are the most common. Enzymes in general improve gas and water retention, stabilize the final product structure, and support consistent quality (Rosell & Collar, 2009). According to the Expert Committee on Food Additives (JECFA), enzymes used as process improvers must demonstrate plausible necessity, be applied following good manufacturing practices, and be dosed at levels which does not affect public health (FAO/WHO, 2006). Additionally, the Codex Alimentarius establishes that food enzymes must avoid misleading the consumer regarding the presence or quality of the food (Codex Alimentarius Commission, 1995). Responsible enzyme use involves dosage control, validation of its performance prior to industrialization, transparency in final consumer declarations, and following the regulatory frameworks at both the international and national levels.

In this context, the present research is structured with two of the objectives from the United Nations Sustainable Development Goals (SDGs) that are directly related to food systems and industrial production (United Nations, 2015). In particular, SDG 2-Zero Hunger, which emphasizes the need to ensure access to safe and nutritious food, and SDG 12-Responsible Consumption and Production, which calls for improving resource efficiency throughout the production chain, reducing waste, and ensuring the correct treatment of hazardous product disposals or use (United Nations, 2015; FAO, 2019). For the type of food treated in this research, bread and tortillas, which depend heavily on flour quality and consistency, improvements at the formulation stages in the laboratory can have a direct influence on food availability, safety, and system quality (FAO, 2018).

By focusing on the analytical evaluation of flour composition, additive use, and process performance, this research contributes to these goals through a preventive and optimization approach rather than reactive quality control. The application of machine-learning methods enables more precise control of chemicals such as enzymes, supports the reduction of variability, and promotes more efficient use of raw materials (FAO, 2017). In doing so, the research aligns technological innovation with sustainability objectives, reinforcing the role of data-driven decision-making as a tool to support responsible production practices within the food industry (United Nations, 2015).

Figure 1- Sustainable Development Goals.

As one of Colombia's longstanding flour and cereal-derived foods producers, Harinera del Valle (HV) has forged a reputation grounded in both tradition and innovation. Founded in the mid-20th century, the company has grown from a single wheat-milling operation into an integrated food group producing wheat and corn flours, industrial and consumer blends, pasta, and edible oils. With a portfolio of more than two dozen brands and a presence across multiple cities and regions, together with recent export activity to several countries, HV is a major player in the Colombian food-processing sector (Harinera del Valle, 2025).

In line with its stated commitment to continual improvement of production, resource efficiency, and environmental responsibility, this research project aims to support HV in the application of machine-learning models to their operational processes. By analysing production, additive-usage, flour composition, and performance data, machine-learning methods can provide HV with insights to optimise enzyme dosing, reduce variability, and enhance the functional performance of flours delivered to bread, tortilla, and other final products.

Despite the widespread use of additives in industrial flour production and in HV, formulation decisions are still largely based on empirical rules, expert judgment, and retrospective analysis of rheological tests. In practice, this means that results are often interpreted only after testing has been completed, limiting the ability to proactively adjust additive dosages to achieve a desired dough behaviour. Given the intrinsic variability of raw materials and the scale of industrial production, this reactive approach can lead to suboptimal use of additives, increased variability in flour performance, and unnecessary reprocessing. The motivation of this research arises from the need to move from descriptive interpretation of dough properties toward predictive and target-oriented formulation support, enabling the R&D team to intentionally design dough profiles that meet specific functional requirements while using additives more efficiently and responsibly.

In response to this motivation, the present study is guided by one central research question:

- To what extent can machine–learning–based, stage-dependent models support formulation and managerial decision-making by reducing uncertainty and reliance on trial-and-error approaches in industrial flour production?

To address this overarching question, the study is structured around the following specific objectives:

1. To assess to what extent machine-learning models can predict dough profiles based on flour characteristics and additive usage under industrial conditions.

2. To identify which stages of the dough rheology test are more reliably predictable using available formulation and compositional variables, and which stages are more sensitive to unobserved factors.
3. To evaluate whether a stage-dependent modelling framework can support targeted formulation by determining additive dosing prior to laboratory testing.

These objectives collectively contribute to developing a data-driven and operationally meaningful framework capable of addressing the central research question focusing not only on model accuracy, but also on understanding the limitations inside the company and the practical value of machine-learning models as decision-support tools in this specific context. In this sense, machine-learning models are positioned not as automated replacements for technical judgment but as analytical tools that enhance managerial oversight, facilitate communication between R&D and production functions, and contribute to more efficient resource allocation in formulation and quality-control activities. Moreover, from a future perspective and to support its broader agenda, the deployment of data-driven analytics complements HV's strategic objectives and helps translate them into measurable operational improvements.

To address this research question and its associated objectives, the study adopts a data-driven methodological approach grounded in the Cross-Industry Standard Process for Data Mining (CRISP-DM). Industrial data generated through routine Mixolab analyses is used to develop stage-dependent machine-learning models that reflect the sequential nature of dough transformation during mixing, heating, and cooling. The analytical framework is designed to align with existing laboratory workflows, using flour characteristics and additive dosages as inputs to predict Mixolab torque parameters across successive stages of the test. Rather than treating dough behaviour as a single outcome, the methodology emphasizes a progressive understanding of how early-stage conditions influence downstream responses. Through this approach, the research aims to deliver a predictive framework capable of supporting targeted dough profiles and more informed additive optimization in industrial flour production.

2. LITERATURE REVIEW

2.1 FLOUR COMPOSITION AND FUNCTIONAL PROPERTIES

Wheat flour is a highly complex material whose functional performance is determined by both its original chemical composition (which could depend on its origin) and the conditions under which it is processed. From a chemical perspective, wheat flour consists primarily of starch (≈65–75%) and proteins (≈8–15%) but also contains lipids, minerals, and minor bioactive components, each of which contributes differently to dough formation during the baking and final product quality (Delcour & Hoseney, 2010; Shewry & Halford, 2002). The interactions between these components form key functional properties such as water absorption, dough elasticity, gas retention, and baking performance.

Starch is the dominant component of flour and plays a central role in determining dough viscosity, temperature behavior, and overall structure. During mixing and baking, starch granules interact with water and proteins, setting the most important characteristics for the dough consistency, which forms the structural matrix of baked products (Shewry & Halford, 2002). Flour proteins are responsible for the elasticity properties of dough. Their compositional and chemical interactions set the dough extensibility and strength, which are critical parameters in differentiating products as bread, pasta or tortillas; as each of those has a different and proper elasticity (Delcour & Hoseney, 2010).

Lipids and enzymes enhance dough behavior through interactions with starch and proteins, influencing softness and overall structure, i.e., crumbs on bread and industrial process tolerance (Goesaert et al., 2005). These now multivariable interactions highlight the intrinsic variability and sensitivity of flour systems.

In addition to composition, flour functionality is strongly affected by previous or original factors such as wheat variety, growing conditions, grain quality, and post-harvest handling, as well as by milling parameters inside each company, including extraction rate, particle size distribution, and blending strategies (FAO, 2019). As a result, flour produced from different origins or samples could exhibit significant variability in results or performance even when nominal composition targets are met, normally starch or protein percentage targets. This variability presents a challenge for flour producers, who must meet the functional requirements of their clients, even using initial input from different sources or changing conditions on the industrial chain, while managing resource efficiency and minimizing waste.

From an industrial perspective, understanding flour composition and its relationship to the results in final products is essential for controlling quality and performance in overall baking processes. Traditional analytical and rheological tests provide valuable insights but often rely on the trial-and-error approach to capture the full complexity of interactions within flour systems, particularly when multiple variables change simultaneously. This complexity in the operational process is the basis for exploring advanced analytical approaches, i.e., machine-learning models, to better predict flour behavior, optimize processing decisions, and support responsible production practices in large operations.

2.2 Role of Enzymes in Flour and Dough Systems

Enzymes play a central role in modern industrialization by allowing modification in dough characteristics through the baking process. Unlike chemical additives, enzymes act as biological catalysts that selectively transform specific attributes within the flour matrix, which influences dough rheology, gas retention, baking performance, and product structure (Van der Maarel et al., 2002). Their widespread adoption in the milling and baking industries reflects their effectiveness in improving product consistency and processing tolerance as scalability enters the overview.

Among the most applied enzymes in flour and dough systems are amylases, xylanases, proteases, and lipases. Amylases support yeast activity, which contributes to loaf volume and crumb softness (Goesaert et al., 2005). Xylanases modify water distribution, improving dough extensibility (Courtin & Delcour, 2002). Proteases act on the gluten proteins, reducing dough resistance and enhancing extensibility, which is particularly relevant for products such as tortillas (Rosell & Collar, 2009). Lipases and phospholipases modify lipid fractions, indirectly strengthening the starch interaction with gluten and proteins, which contribute to crumb structure and life extension (Van der Maarel et al., 2002).

Despite their benefits, enzymatic effects are highly dependent on flour composition, processing conditions, and dosage levels. Errors in the application can result in undesirable outcomes, such as products with weak structure or waste increase (Goesaert et al., 2005). Additionally, enzyme usage and outcome are influenced by other factors, including pH, temperature, and mixing time, which vary across production environments and flour samples. This sensitivity highlights the complexity of enzyme interactions with the flour and the limitations of fixed dosing strategies in industrial contexts.

From a regulatory and safety view, enzymes used in food processing are classified as “processing aids” and must comply with strict international guidelines. Views already stated from JECFA, Codex Alimentarius, or other entities/regulations reinforce the importance of responsible enzyme use, not only from a quality perspective but also in terms of regulatory compliance and consumer trust.

In industrial milling operations such as those studied in this research, the challenge therefore lies in balancing enzymatic functionality with process robustness, product consistency, and sustainability objectives. Given the scale of flour production and the variability inherent in raw materials, traditional trial and error approaches to enzyme selection and dosing are insufficient or simply resource and time-consuming. This context provides a strong rationale for the application of advanced analytical tools capable of integrating compositional, processing, and performance data. In this regard, machine-learning approaches offer a promising pathway to model complex enzyme–flour interactions, support optimized enzyme use, and reduce waste and variability in large-scale flour production systems.

2.3 Sustainability and Responsible Production in Flour Industries

The sustainability of flour and baking industries has gained increasing attention as part of a big picture transformation of global food systems toward more responsible consumption and production patterns. Cereal-based foods represent a major share of daily caloric intake worldwide, making the environmental and resource efficiency of flour production particularly relevant in achieving international sustainability objectives such as the SDGs (FAO, 2018).

Flour production is an energy and resource-intensive activity, involving the grain cleaning, milling, blending, and overall formulation processes that all together contribute to gas emissions, water use, and final product generation, which itself includes other waste sources, such as the plastic for the packaging. According to FAO assessments, a substantial share of food losses in cereal supply chains occurs at post-harvest, processing, and baking stages, underscoring the importance of improving efficiency at these early transformation points (FAO, 2019). Inadequate process control, variability in the origin of raw materials, and suboptimal use of, for example, enzymes can increase losses and lead to unnecessary waste, which affects both economic performance and sustainability goals.

Responsible production in the flour and baking sectors goes beyond grain and raw material selection and attention. It is imperative to also include careful use and management of additives such as enzymes. SDG 12 explicitly has two targets which could encompass the use of these enzymes, Responsible Management of Chemicals and Waste (Target 12.4), as well as Sustainable Management and Use of Natural Resources (Target 12.2) (United Nations, 2015). In this context, the responsible use of enzymes, applied to specific flour characteristics inside the flour industry, supports both regulatory compliance and sustainability objectives by reducing unnecessary chemical waste, minimizing reprocessing, and improving industrial consistency.

Industry initiatives at all levels demonstrate the growing emphasis on sustainability within the cereal and baking sectors. Programs such as the Sustainable Wheat Initiative and similar approaches promoted by associations on the national or international level emphasize optimization of resource use and improved final quality of the products. These initiatives recognize that achieving sustainability improvements requires not only changes in agricultural practices or movements across the production chain but also innovation at the formulation stage, where data-driven optimization can play a critical role. (FAO, 2019; United Nations, 2015)

Within this framework, integrating advanced analytical and predictive tools offers significant potential to support responsible production. By enabling early detection of quality deviations, more precise control of processing aids, and improved alignment between flour characteristics and end-product requirements, digital and data-driven approaches contribute to waste prevention rather than post-process corrections (FAO, 2017; Zhang et al., 2025). As such, sustainability in flour production is increasingly understood not as a separate objective but as an outcome of improved process understanding, optimization, and decision-making within industrial operations (FAO, 2019; UNEP, 2021).

2.4 Machine Learning in Food Processing and Quality Control

The increasing complexity of food industrial systems, combined with growing sustainability and efficiency demands, has driven interest in advanced data-driven methods for monitoring, prediction, and optimization inside industrial production (FAO, 2017; Kourti & MacGregor, 1995). Machine learning (ML), a subset of artificial intelligence, has emerged as a valuable tool due to its capacity to model nonlinear relationships, handle multivariate datasets, and extract patterns from large volumes of production data (Zhang et al., 2025, Costa, 2025). In contrast to traditional statistical approaches, which often rely on predefined assumptions and limited variable interactions, ML methods are particularly well-suited to capturing the complex interdependencies characteristic of food systems such as the flour baking process (Pathare & Roskilly, 2016).

Within the food industry, machine-learning techniques have been applied across a wide range of domains, including quality prediction, life estimation, process control, defect detection, and formulation optimization (Zhang et al., 2025; Vásquez-Caicedo et al., 2018; Pathare et al., 2013). Supervised learning algorithms such as artificial neural networks (ANN), support vector machines (SVM), and random forest models have demonstrated strong performance in predicting quality attributes based on raw material properties and processing conditions (Pathare & Roskilly, 2016; Zhang et al., 2025). These approaches are especially relevant in cereal processing, where flour composition, additive usage, and processing parameters interact in highly nonlinear and mostly interdependent ways.

In flour milling and baking applications, ML has been explored to predict rheological properties of the flour, baking performance, and product quality based on compositional and process data that have been studied in niche universities. Studies have shown that data-driven models can outperform traditional regression methods when dealing with heterogeneous flour batches and variable process conditions, offering improved robustness and predictive accuracy, at least on the laboratory applications available (Kotsiantis et al., 2007). Importantly, ML models can integrate data from routine industrial measurements, enabling their deployment at scale without the need for costly or time-intensive experimental testing.

Despite these advances, the application of machine learning in flour production remains relatively limited compared to other food sectors, and many published studies are restricted to laboratory-scale datasets or simplified experimental conditions. There is a notable gap in industrial-scale implementations that integrate real production data, particularly in the context of enzyme optimization and sustainable production objectives. Addressing this gap requires applied research that combines domain knowledge in cereal science with robust ML methodologies, validated under real manufacturing conditions. This study seeks to contribute to this emerging field by applying machine-learning models to industrial flour-production data, with the dual aim of improving process performance and supporting responsible production practices.

2.5 Theoretical Implications

The existing body of literature provides a robust foundation regarding flour composition, enzyme functionality, sustainability challenges in cereal processing, and the growing application of machine-learning techniques in food systems. However, when these domains are considered collectively, several critical research gaps emerge that justify the present study.

First, although the role and effects of enzymes in flour are well documented, most studies focus on isolated enzymatic effects under controlled laboratory conditions or fixed flour compositions. These approaches cannot capture the variability in industrial flour production, where grain heterogeneity and changing process conditions influence enzyme performance. There remains limited empirical research addressing how enzymatic treatments interact with variable flour compositions at an industrial scale, particularly in continuous production environments.

Second, while sustainability and responsible production are widely discussed in the context of cereal supply chains, relatively few studies translate these objectives into actionable process-level decision tools within the flour industry, and moreover, their formulation operations. Existing sustainability frameworks used in similar research frequently emphasize agricultural practices or waste reduction, with less attention given to optimization opportunities during flour processing and additive management. Consequently, there is a gap between sustainability targets and the operational mechanisms required to achieve them in large-scale flour production, i.e., formulation inside the baking process.

Third, despite the demonstrated potential of machine-learning methods in food engineering, their application within flour milling and enzyme optimization remains limited. There is a lack of studies integrating multivariable production data, including flour composition, enzyme usage, and performance outcomes, to support predictive and preventive process control strategies aligned with sustainability objectives.

Finally, there is a notable scarcity of applied case studies conducted in Latin American industrial contexts. Given the regional importance of foods based on cereals and the scale of flour consumption, the absence of data-driven research based on local production systems represents a missed opportunity for both academic and industrial impact. Addressing this gap is especially relevant for countries such as Colombia, where flour producers play a critical role in national GDP and, by extension, the agro-industrial outlook.

To tackle these challenges, the present study adopts and develops the Cross-Industry Standard Process for Data Mining (CRISP-DM) as its methodological framework. CRISP-DM as a method provides a structured and learning iterative approach, organizing the process into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Chapman, 2000). Its flexibility, aligning analytical tasks with real-world objectives, permits its continued adoption in industrial and applied research contexts, including manufacturing applications, as the one being performed, specially used in a context of design science approach (Plotnikova, Dumas & Milani, 2020, Aparicio et. al, 2023).

3. Managerial Implications

Beyond its technical contributions, this research has important managerial implications for how formulation processes, decision-making structures, and cross-functional coordination are organized within industrial flour production environments.

A central implication of this study is the potential to reconfigure the formulation and commercial negotiation workflow. In the current process, commercial agreements are often established based on predefined flour profiles, while technical feasibility and cost implications are assessed later through iterative laboratory testing. This sequencing introduces uncertainty and shifts operational risk toward R&D and production teams. The modelling framework developed in this research enables an alternative approach in which technical feasibility can be evaluated prior to commercial commitment, by combining flour characteristics with target dough profiles to generate initial additive recommendations. This supports earlier alignment between customer requirements, technical feasibility, and cost constraints.

From an operational viewpoint, this acts as a transition from reactive to proactive formulation. Instead of relying on trial-and-error approaches, R&D teams or even other technical teams inside the company can use insights from the models to design formulations that are closer to desirable outcomes from the start. This shift can reduce the number of iterations required, shorten negotiation cycles, and improve resource employment. Such transformations are consistent with evidence in the literature showing that the use of predictive analytics in manufacturing or industrial settings can significantly improve process performance and reduce variability (Chui et al., 2018).

The framework also sought to strength cross-functional integration within the organization. By providing a shared analytical basis for evaluating formulation feasibility, the model facilitates communication between commercial, R&D, production, and finance teams. Commercial teams acquire better visibility into technical constraints, R&D teams can quantify the impact of formulation adjustments, and finance teams can better assess cost implications associated with additive usage and finished goods production. This type of data-driven coordination is a defining characteristic of organizations that successfully leverage analytics to improve decision-making and operational alignment (McAfee & Brynjolfsson, 2012).

In addition, the potential integration of model outputs into operational systems such as ERP platforms supports improved traceability and governance. By linking formulation recommendations with actual production outcomes and material consumption, the company can create a structured feedback loop that enables continuous learning and performance monitoring. This contributes to more controlled and auditable processes, which are particularly relevant in industrial food production environments.

Finally, the proposed approach contributes to more responsible production practices. By optimizing dosing and reducing unnecessary experimentation, the framework supports more efficient use of resources as it diminishes the number of iterations to obtain a desirable outcome. This aligns with the growing emphasis on integrating optimization into industrial production, where improved process control is recognized as a key driver of both economic and environmental performance (Bertsimas & Kallus, 2020).

Overall, the managerial contribution of this research lies in demonstrating how machine-learning models can be embedded within organizational processes to support more informed decision-making, reduce uncertainty, and improve coordination across functions. Rather than replacing expert judgment, the framework enhances it by providing a structured and quantitative foundation for formulation and production decisions.

4. METHODOLOGY

The methodological approach in which this research is designed, not only aims to address a technical modelling problem, but also to respond to a managerial challenge within industrial flour production. Specifically, the objective is to develop an analytical model capable of supporting formulation and decision-making processes under variable conditions of raw-material, cost constraints, and commercial commitments. From a managerial perspective, this requires tools that connect team needs and at the same time enable earlier and more informed decisions regarding product negotiations with clients and formulation strategies. The methodology prioritizes alignment between analytical development and operational realities, ensuring that model outputs can be interpreted, communicated, and used by R&D also known as the technical team and C-level stakeholders rather than remaining confined to purely statistical evaluation.

In this study, CRISP-DM is not treated as a rigid sequence of technical steps, but rather as a conceptual framework that guides how business objectives, data analysis, and modelling decisions are connected. The methodology is designed to ensure coherence between the industrial problem and the analytical solutions developed (Chapman et a., 2000, Costa & Aparicio, 2020,).

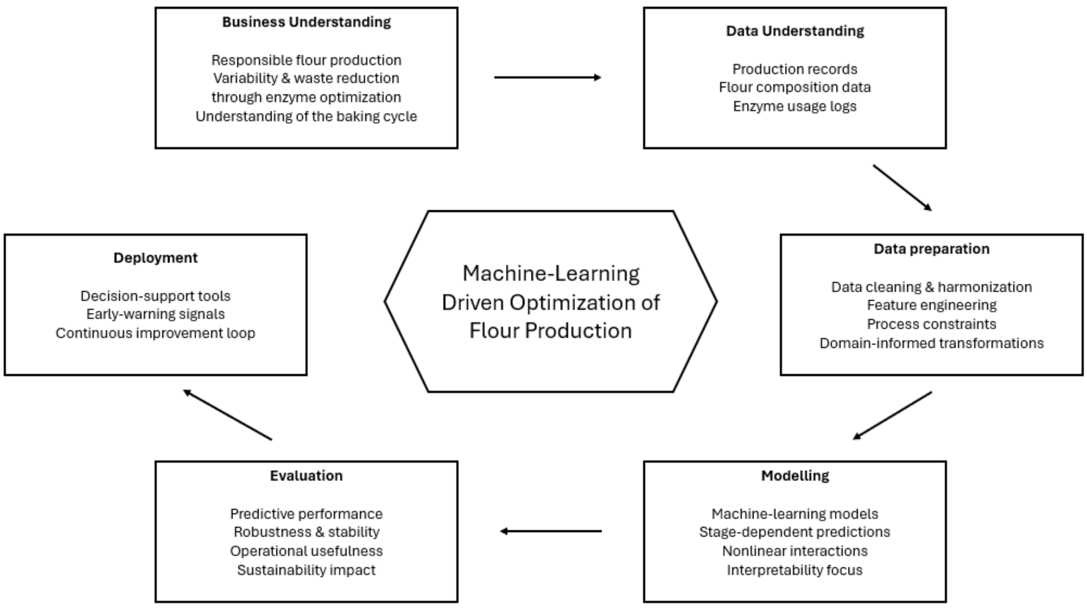

Figure 2- CRISP-DM Methodology applied to the research

4.1 Business Understanding

The business understanding activity defines the industrial context in which this research is conducted and clarifies how analytical objectives arise directly from routine production and quality-evaluation practices.

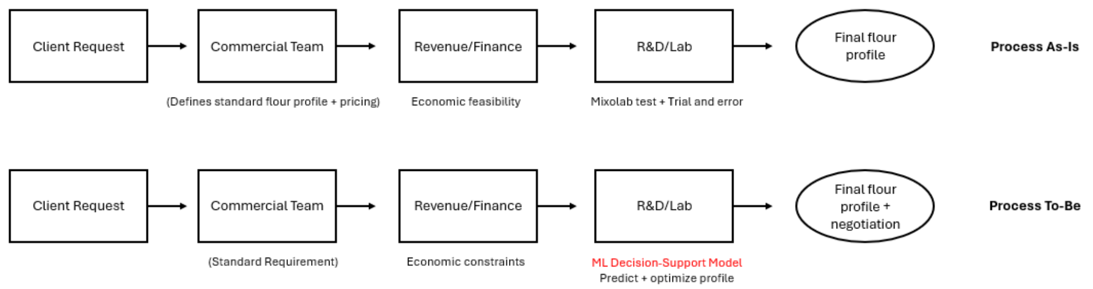

Figure 3- Flour profile negotiation process

In the current operational flow at Harinera del Valle, commercial teams act as the first interface with clients, defining commercial agreements and committing to the supply of flour based on a set of standard or historically validated functional profiles. Economic conditions and margin considerations are addressed jointly by commercial and revenue management teams, ensuring financial feasibility prior to production commitment. Once these agreements are in place, the responsibility shifts to the R&D team, in coordination with finance and production, to translate commercial promises into a viable flour formulation specially made for the specific client. At this stage, productivity, additive usage, and flour performance must be evaluated to ensure alignment between functional requirements and cost efficiency. However, this process needs to shift its order, as profit measurement and profile definition should happen before commercial deals are closed, this is impossible as is still largely supported by empirical adjustments and iterative laboratory testing.

At Harinera del Valle (HV), flour quality and functional performance are measured on a day-to-day basis using the Mixolab system, a standardized rheological analysis tool widely adopted in the milling and baking industries. The Mixolab provides information on dough mixing behaviour, protein weakening, starch gelatinization, and enzymatic activity as a function of time and temperature, thereby offering an integrated view of flour functionality under controlled conditions.

From an operational standpoint, the Mixolab serves as a central reference for formulation decisions, enzyme adjustments, and quality validation. The measurements generated by this system are not isolated laboratory observations but rather form part of the routine feedback loop between production, quality control, and formulation teams, i.e., at HV, the R&D team runs the formulation and Mixolab while working closely with production. As such, the data collected reflect both the inherent variability of raw materials and the cumulative impact of process and formulation choices made upstream.

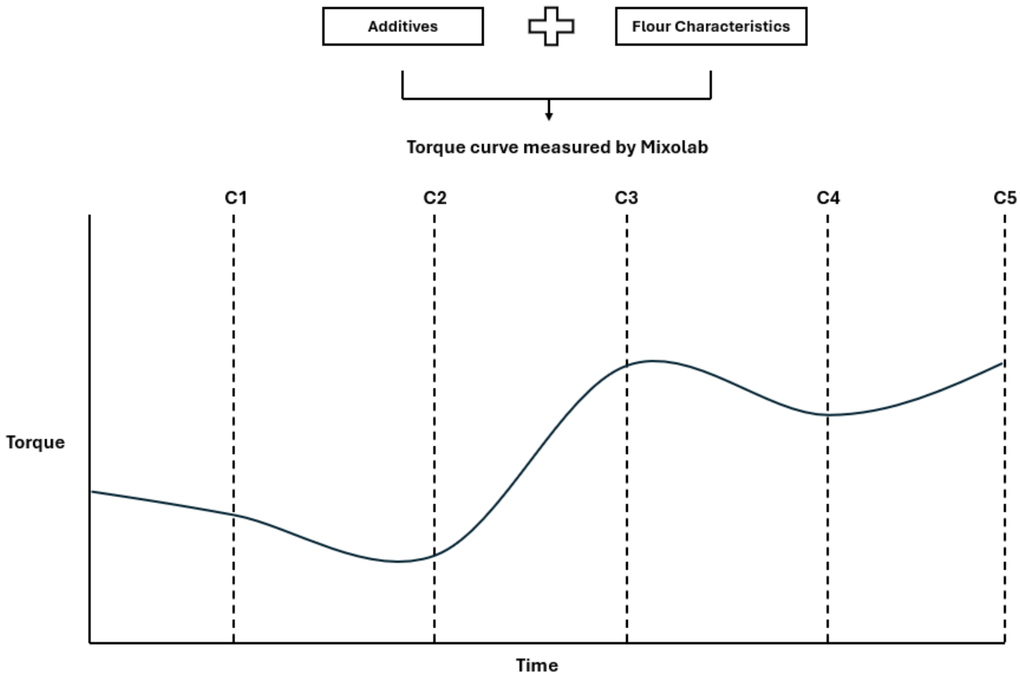

Figure 4 - Mixolab results curve - Change of torque across time

A defining characteristic of Mixolab data is its stage-based and sequential structure. The test protocol progresses through a series of phases that correspond to distinct physical and biochemical phenomena occurring in dough systems, with each phase influenced by the conditions established in preceding ones. Consequently, parameters derived from early stages constrain and shape the behaviour observed in later stages. This sequential dependency mirrors the logic of industrial flour production, where early formulation and process decisions propagate through the system and influence final performance.

The test begins with the initial mixing phase (C1), during which flour hydration, protein network development, and mechanical resistance are established under constant temperature. The torque measured at this stage reflects the ability of the flour to form a stable dough, while the associated stability time indicates the resistance of the dough to mechanical and thermal stress.

As the temperature increases, the dough enters the protein weakening phase (C2). In this stage, thermal and mechanical stresses lead to partial disruption of the protein network, resulting in a reduction of dough consistency. The minimum torque observed during this phase provides information on protein strength and thermal tolerance.

Further heating leads to the starch gelatinization phase (C3), characterized by a sharp increase in torque as starch granules swell and gelatinize in the presence of water. Parameters associated with this stage describe the capacity of the flour to develop viscosity and structure under thermal conditions.

At higher temperatures, the dough enters the enzymatic degradation phase (C4), during which starch polymers lead to a decrease in dough consistency. Torque measurements at this stage provide insight into the activity of enzymes with the starch matrix.

Finally, during the cooling phase, the dough passes through starch retrogradation (C5). The increase in consistency observed at this stage reflects the new links of starch chains and the formation of a gel structure. Parameters derived from this phase are often associated with textural properties and structural stability in finished products.

The figure highlights how formulation inputs propagate through these sequential stages, shaping the overall torque profile across time, and provides the basis for stage-dependent analysis and targeted dough profile optimization.

In addition to torque values, the dataset includes slope parameters describing the rate of change between stages, as well as time and temperature descriptors indicating when specific transitions occur during the test. Together, these variables provide both static and dynamic representations of dough behaviour.

The central business challenge addressed in this study arises from the difficulty of translating this rich, multistage information into predictive and preventive decision support. While Mixolab outputs provide valuable descriptive indicators, their interpretation is often retrospective and relies heavily on expert judgment. In practice, this limits the ability to anticipate downstream behaviour based on early-stage signals or to quantitatively assess how adjustments in enzyme dosage or formulation parameters will affect later stages of the process.

Accordingly, the primary business objective of this research is to enhance the value of existing Mixolab data by enabling stage-aware, data-driven predictions that reflect the sequential nature of flour behaviour. Rather than treating Mixolab parameters as independent or static descriptors, the study aims to model their interdependencies in a manner consistent with the progression of the Mixolab test. This objective directly motivates the adoption of analytical strategies in which predictions at one stage inform subsequent stages, aligning the modeling approach with both the structure of the data and the operational logic used by technical teams.

4.2 Data Understanding

The data used in this study is derived entirely from routine Mixolab analyses conducted by Harinera del Valle as part of its standard flour evaluation workflow. The variables available in the dataset reflect both formulation inputs and stage-specific functional responses, providing a comprehensive representation of flour performance under standardized conditions.

Table 1- Controllable Variables Description

Variable group	Variable name	Description	Role in analysis
Additives – Initial features that have incidence across all the process C1-C5	amylase	Dosage level of amylase enzymes added to the flour formulation. Influences starch hydrolysis, fermentable sugar availability, and dough viscosity during heating.	Independent variable; controllable input
	xylanase	Dosage level of xylanase enzymes applied to modify arabinoxylans and improve water distribution and dough extensibility.	Independent variable; controllable input
	protease	Dosage level of protease enzymes used to modify gluten protein structure and reduce dough resistance.	Independent variable; controllable input
	lipase	Dosage level of lipase or phospholipase enzymes affect lipid interactions and crumb structure development.	Independent variable; controllable input

Table 2- Non-Controllable Variables Description

Flour characteristics – Initial attributes which vary across flour batches, have incidence across all the process C1-C5	protein	Total protein content of the flour (%). Influences dough strength, water absorption and protein network formation.	Conditioning variable; structural indicator
	ash	Ash content (%), reflecting mineral content and flour extraction rate.	Conditioning variable; structural indicator
	moisture	Moisture content of the flour prior to testing (%). Affects effective water availability and dough consistency.	Conditioning variable; structural indicator
	damaged_starch	Proportion of mechanically damaged starch granules generated during milling. Influences water absorption and enzymatic susceptibility.	Conditioning variable; structural indicator
	falling_number	Indicator of endogenous amylase activity in the flour. Lower values indicate higher enzymatic activity.	Conditioning variable; structural indicator
Global flour property measured in C1	absorption	Water absorption determined during the initial phase of the Mixolab test. Reflects the amount of water required to reach target dough consistency and conditions all subsequent dough behavior.	Key conditioning variable; early-stage predictor
Torque in C1, C2, C3, C4 and C5	C1,C2,C3,C4,C5	Dough consistency (torque) during each stage of the mixing.	Early-stage target and conditioning variable
Stability measured between C1 and C2	time_stab	Duration for which dough maintains stable consistency during the C1 phase. Indicates resistance to mechanical and thermal stress.	Constraint variable; gating condition
Slope parameters, changes between torques, measured between C1 and C4	slope_alpha	Rate of change in torque between C1 and C2, describing protein weakening dynamics.	Intermediate predictor
	slope_beta	Rate of change in torque between C2 and C3, associated with starch gelatinization kinetics.	Intermediate predictor
	slope_gamma	Rate of change in torque between C3 and C4, associated with the enzymatic starch degradation rate.	Intermediate predictor

At a high level, the dataset can be grouped into three main categories: formulation and additive variables, global flour properties and stage-dependent Mixolab parameters. This structure reflects the operational logic of the Mixolab test and the way results are interpreted in industrial practice.

The formulation-related variables include information on flour composition and additive usage, such as enzyme-related inputs and other formulation components applied to the flour prior to testing. The enzymes represent controllable decision levers from a production perspective and are treated as independent inputs across all analytical stages, while the flour composition is a given characteristic of our raw materials. In addition, the dataset includes water absorption, which is a global flour property determined during the initial phase of the Mixolab test. Absorption is a critical variable, as it influences dough consistency throughout the entire test and constrains the behaviour observed in subsequent stages.

The core of the dataset consists of stage-dependent Mixolab parameters, which describe dough torque evolution across the standard Mixolab phases. These parameters include torque values associated with distinct stages of the test (commonly denoted as C1, C2, C3, C4, and C5), each corresponding to specific physicochemical phenomena such as protein development, protein weakening, starch gelatinization, and enzymatic degradation. In addition to torque values, the dataset contains slope parameters (e.g., α, β, γ), which quantify the rate of change between stages and provide insight into the dynamics of dough transformation rather than static end-point behaviour.

A variable of relevance is the stability time, which characterizes the resistance of the dough to mechanical and thermal stress during the initial mixing phase. This variable serves as an important constraint within the dataset, as it defines minimum performance thresholds that must be satisfied before downstream stages are considered meaningful.

4.3 Data Preparation

The data preparation activity focuses on transforming raw industrial data generated through Mixolab analyses into a consistent and reliable dataset for subsequent analysis. Given that the data originates from routine production and quality-control activities, this phase prioritizes data integrity, physical plausibility, and consistency with the operational logic of the Mixolab test.

The initial step in data preparation is plain data cleaning and validation. All observations are screened for completeness and internal consistency, with particular attention to key formulation inputs, flour characteristics, and Mixolab parameters. Observations with missing or inconsistent values in critical variables are handled selectively, ensuring that incomplete records do not introduce bias or distort the analytical dataset. Where missing values reflect incomplete test execution rather than measurement error, affected records are excluded from further analysis to preserve the reliability of the dataset.

A second key component of data preparation is the application of physical and procedural constraints. All quantitative variables are constrained to non-negative values, reflecting the physical reality of formulation inputs, compositional measurements, and rheological parameters. Additionally, constraints inherent to the Mixolab methodology are explicitly enforced. Stability time is treated as a prerequisite condition, and observations that do not meet a minimum stability requirement are excluded from further consideration. This ensures that downstream Mixolab parameters are only analyzed when the initial dough development phase has reached a meaningful level of stability.

Data transformation during this phase is guided by process coherence rather than statistical convenience. Variables are organized by role within the Mixolab test sequence, preserving the logical progression from initial mixing to later thermal stages. This organization ensures that the prepared dataset reflects the temporal structure of the test and avoids mixing information from non-comparable phases.

4.4 MODELLING

The modelling activity focuses on the development of machine-learning models capable of capturing the sequential and interdependent nature of Mixolab-derived flour performance data. In contrast to approaches that treat all response variables as independent or simultaneous targets, this study adopts a stage-dependent modelling strategy that reflects the natural progression of the Mixolab test and the cumulative character of dough transformation. The final dataset used for modelling consisted of 125 valid Mixolab observations after applying data-cleaning and stability constraints described in Section 4.3. All models were trained using a consistent split of 100 samples for training and 25 samples for holdout testing, ensuring comparability across Mixolab stages and derived parameters. A total of ten stage-dependent targets were modelled, following the chronological order of the Mixolab test:

Table 3 - Description of Stage-Dependent Models

Model	Mixolab stage / parameter	Dependent variable (target)	Independent variables (predictors)
1	Absorption (C1)	Water absorption at C1	Flour characteristics; additive dosages
2	Initial mixing (C1)	Torque at C1	Flour characteristics; additive dosages; predicted absorption
3	Stability (CS)	Torque at CS	Flour characteristics; additive dosages; predicted absorption; predicted C1 torque
4	Protein weakening rate	Slope α (C1–C2)	Flour characteristics; additive dosages; predicted C1 and CS outputs
5	Protein weakening minimum (C2)	Torque at C2	Flour characteristics; additive dosages; predicted C1, CS and slope α
6	Gelatinization rate	Slope β (C2–C3)	Flour characteristics; additive dosages; predicted outputs up to C2
7	Gelatinization peak (C3)	Torque at C3	Flour characteristics; additive dosages; predicted outputs up to slope β
8	Enzymatic degradation rate	Slope γ (C3–C4)	Flour characteristics; additive dosages; predicted outputs up to C3
9	Enzymatic degradation (C4)	Torque at C4	Flour characteristics; additive dosages; predicted outputs up to slope γ
10	Cooling / retrogradation (C5)	Torque at C5	Flour characteristics; additive dosages; predicted outputs from all prior stages

This ordering reflects the progressive and cumulative nature of dough transformation and was preserved throughout training, evaluation, and inverse analysis. The analytical framework is built around a series of single-target regression models, each corresponding to a specific Mixolab stage or derived parameter. Rather than fitting a single global model to predict all outcomes simultaneously, individual models are trained for each stage in the order in which they occur during the test. This choice is motivated by both process logic and interpretability considerations. From a process perspective, later-stage behaviour cannot occur independently of earlier stages; from an operational perspective, intermediate predictions are meaningful and actionable.

Random Forest regression is selected as the core modelling technique due to its ability to model nonlinear relationships, capture complex interactions among heterogeneous variables, and maintain robustness when applied to noisy industrial datasets (Breiman, 2001; Hastie, Tibshirani, & Friedman, 2009). These properties are particularly relevant in food-processing systems, where functional responses arise from multivariate and non-additive interactions between formulation inputs, flour characteristics, and process conditions. In addition, Random Forest models are relatively insensitive to variable scaling, can handle correlated predictors effectively, and are less prone to overfitting than many high-capacity models when properly tuned, making them well-suited for industrial applications with moderate sample sizes.

Beyond predictive performance, Random Forest models offer practical advantages in terms of interpretability and stability, which are important for adoption in this specific operational context. Variable-importance measures and partial-dependence analyses allow technical teams to identify influential factors and understand model behaviour without relying on opaque representations. This balance between accuracy and interpretability has been highlighted in the literature as a key reason for the widespread adoption of Random Forest methods in applied food science and process engineering contexts (Zhang et al., 2025).

Then becomes important to assess the performance of the developed models with the objective of understanding how predictability varies across the different stages of the Mixolab test. Rather than aiming solely to maximize global predictive accuracy, it is designed to identify which stages of dough development can be reliably predicted based on the available formulation, compositional, and process variables, and which stages exhibit greater uncertainty potentially associated with unobserved or external factors.

Table four summarizes the predictive performance of the Random Forest models for each target on the holdout test set, using R^2, MAE, and percentage MAE (%MAE) as indicators of plausibility and robustness.

Table 4 - Predictive performance by Mixolab target using Random Forest Regressor

Target	R^2	MAE	%MAE
Water absorption at C1	0.847	0.0025	0.42
Torque at C1	0.997	0.0007	0.06
Torque at CS	0.971	0.0024	0.23
Slope α (C1–C2)	0.996	0.0005	0.41
Torque at C2	0.959	0.0060	1.39
Slope β (C2–C3)	0.993	0.0036	0.98
Torque at C3	0.977	0.0063	0.38
Slope γ (C3–C4)	-1.727	0.0315	68.98
Torque at C4	0.992	0.0062	0.39
Torque at C5	0.976	0.0110	0.44

The results show strong predictive performance for most torque-related stages, particularly C1, CS, C3, C4, and C5, with %MAE values consistently below 1%. These stages correspond to dough consistency levels that are strongly conditioned by flour characteristics and additive dosing, which are explicitly represented in the dataset.

Slope parameters α and β also exhibit excellent predictability, indicating that rates of structural change during protein weakening and starch gelatinization can be reliably captured using the available formulation and compositional variables.

In contrast, slope γ (C3–C4) shows poor predictive performance, with negative R^2 and high relative error. This result suggests that the enzymatic degradation dynamics captured in this stage are likely influenced by factors not fully represented in the dataset.

When analyzed along the Mixolab sequence, the results reveal a clear pattern:

- Early and mid-stages (C1–C3) exhibit very high predictability.
- Late torque stages (C4, C5) remain predictable despite accumulated process variability.
- Transition slopes involving enzymatic degradation (γ) present the highest uncertainty.

This stage-wise differentiation supports the methodological assumption that not all parts of the Mixolab curve are equally predictable using standard industrial measurements, and that some stages are inherently more sensitive to unobserved factors.

On top of forward prediction, the modelling framework is designed to support a target-oriented or inverse analytical perspective. Desired Mixolab torque values or stage profiles are treated as target conditions, and the trained models are used to explore how variations in additive dosing influence the ability to reach those targets. Rather than directly solving an optimization problem, this inverse use of the predictive models enables iterative evaluation of additive configurations that minimize deviation from the desired torque profile. This approach aligns with industrial formulation practices, where acceptable operating ranges and practical constraints guide decision-making, and supports the use of machine-learning models as formulation decision-support tools rather than black-box optimizers.

The trained models were used in a target-oriented inverse setting, where desired Mixolab curves were treated as objectives and additive dosages were iteratively adjusted to minimize deviation from those targets. From a business perspective, this capability is particularly relevant as it opens the possibility of rethinking the current negotiation and formulation workflow: instead of committing to standard flour profiles and subsequently adapting formulations through trial-and-error, the company could preliminarily evaluate whether a requested dough profile is technically and economically feasible before commercial commitments are finalized.

Inverse optimization was evaluated on ten randomly selected production samples, covering a wide range of flour characteristics and additive configurations. Across these cases, the optimization procedure consistently produced small, bounded adjustments in additive dosages, rather than extreme or unrealistic changes. Table five summarizes the outcomes.

Table 5 - Summary of Inverse Optimization Outcomes Across 10 Samples

Metric	Minimum	Maximum	Mean	Interpretation
Best optimization loss	0.0018	1.7149	0.456	Indicates generally good convergence, with higher loss in stages affected by enzymatic degradation, as the slopes.
Number of additives adjusted	2	6	4.3	Optimization selectively adjusts only relevant additives per formulation
Maximum additive change (ppm)	0.02	12.20	4.1	Adjustments remain within realistic industrial ranges.
Mean absolute change per additive (ppm)	0.21	6.06	2.3	Confirms optimization, refining dosing rather than radically altering formulations
Max absolute torque error (Nm)	0.00003	0.195	0.024	Larger deviations were observed mainly in late or enzymatic-driven stages
Mean torque error – early stages (C1–C3)	< 0.001	0.004	0.0016	Early and mid-stages are consistently well controlled.
Mean torque error – late stages (C4–C5)	0.0001	0.195	0.037	Higher variability due to accumulated uncertainty
Cases with improvement vs real additives	8	10	9.5	Optimization improves curve matching in most scenarios
Unrealistic additive recommendations	0	0	0	Confirms operational feasibility of optimization results

Across all cases, optimization converged with bounded losses and produced moderate, realistic adjustments in additive dosing. The largest improvements were consistently observed in early and mid-Mixolab stages (C1–C3), while higher residual errors appeared in stages associated with enzymatic degradation and cooling (meaning the latest stages in which many effects added up across time). Importantly, the optimization process did not generate extreme or impractical additive recommendations, reinforcing its suitability as a formulation decision-support tool rather than an automated control system.

In addition to Random Forest models, gradient boosting–based approaches were also explored during the modelling phase, including tree-based boosting algorithms commonly used for regression tasks. Although these methods are known for their strong predictive performance in many domains (Friedman, 2001), their application in this study did not yield consistent improvements over Random Forest models. Boosting approaches exhibited greater sensitivity to noise and data heterogeneity, leading to less stable performance across stages and increased risk of overfitting under industrial data conditions. Moreover, their sequential error-correction mechanism proved less compatible with the already sequential structure imposed by the Mixolab stages, complicating interpretation and reducing transparency of intermediate predictions.

Given these considerations, Random Forest regression was retained as the primary modelling approach, as it provided a more robust and interpretable balance between predictive accuracy, stability, and operational relevance. To further ensure physical plausibility, constraints are applied to model outputs, with predicted values restricted to non-negative ranges consistent with the nature of rheological and compositional parameters. This prevents the generation of unrealistic predictions and reinforces alignment with real production conditions.

Overall, the model was designed to produce a process-aware and operationally meaningful analytical framework rather than a purely technical predictive solution. By combining stage-specific Random Forest models with chained predictions and a target-oriented perspective, the approach aligns machine-learning techniques with the physical structure of the Mixolab test and the formulation decision-making needs of industrial flour production.

4.5 Evaluation

The evaluation activity in this research is framed as an ongoing and progressive collaboration with Harinera del Valle rather than as a finalized implementation step. While full operational integration of the modelling framework is beyond the scope of the present study, the work has been conducted in close interaction with technical teams to assess the practical usefulness and optimization capability of the proposed approach.

Rather than focusing solely on predictive performance, current efforts emphasize evaluating the decision-support value of the models for additive optimization. In collaboration with formulation and R&D teams, the models are used to explore alternative additive-dosing scenarios and to assess their ability to guide formulations toward desired Mixolab torque profiles. This process allows the company to measure how effectively the models support formulation choices, reduce trial-and-error experimentation, and improve alignment between target dough profiles and laboratory outcomes.

A staged validation strategy is being followed. In the first stage, the models are applied to predict Mixolab torque curves prior to laboratory testing, using known flour characteristics and additive inputs. Predicted profiles are then compared against experimental Mixolab results, enabling direct assessment of the models' ability to approximate targeted rheological behaviour. This phase supports internal evaluation of model reliability and sensitivity to formulation changes under controlled conditions.

Subsequent activities focus on extending this evaluation toward industrial relevance, examining how model-guided formulation adjustments translate into process consistency and downstream performance. This step is not intended as immediate automation, but rather as a structured learning phase in which model outputs, expert judgment, and laboratory results are jointly assessed to refine both the analytical framework and formulation practices.

Throughout this process, emphasis is placed on iterative connected work, interpretability, and the share of knowledge between teams. Feedback from R&D is used to evaluate the robustness of optimization recommendations and to identify areas where additional data or constraints may improve model performance. This collaborative and incremental approach supports gradual adoption into the production environment while ensuring that the analytical framework remains grounded in operational reality.

In summary, evaluation in this research is understood as a continuous validation and learning process, aimed at measuring and strengthening the optimization power of machine-learning models within industrial flour formulation. Ongoing work with Harinera del Valle seeks to transition the framework from an analytical tool toward a practical decision-support system capable of enhancing additive optimization and supporting responsible production practices over time.

To make the evaluation concept tangible for Harinera del Valle, an inverse-design use case was tested on a real historical Mixolab observation selected from the evaluation set used to train the model. In this scenario, the flour characteristics of the batch are treated as fixed (the raw material condition that production has available), while the desired outcome is defined as the target Mixolab profile that the company wants to deliver to a specific client. The objective is therefore not to "predict what will happen" but to determine what additive configuration would best reproduce the desired torque profile under the same flour conditions.

Using the trained stage-dependent models, additive dosages were iteratively adjusted within realistic bounds to minimize the deviation between the predicted Mixolab parameters and the target profile. This produces a practical recommendation: given this flour batch and the requested rheological target, the model proposes enzyme levels that should achieve the closest match without relying on repeated trial-and-error Mixolab runs. In this example, the optimization converged with a very low final loss (0.00177), and the recommended changes remained moderate.

Table 6 - Inverse-design deployment example: recommended additive dosing to reach a target Mixolab profile

Additive (ppm)	**Real dosing**	**Model-recommended dosing**	**Δ (recommended − real)**
AA	38.78	40.48	+1.70
GOX	5.10	5.81	+0.71
LI	16.96	19.81	+2.85
XB	0.00	0.02	+0.02
XF	0.00	0.00	0.00
Aa	30.39	30.18	−0.21

This type of result illustrates how the model could support a revised workflow: instead of finalizing customer commitments based on standard profiles and then adjusting formulation through iterative tests, the organization could pre-evaluate feasibility by combining (i) the flour batch characteristics available in production and (ii) the client's target dough profile, generating an initial additive recommendation for R&D validation. In practice, this would reduce formulation uncertainty, shorten development cycles, and strengthen coordination between commercial commitments and technical delivery. The full implementation is publicly available (Marin, 2026).

4.6 Deployment

The deployment phase of this research is framed as a forward-looking integration roadmap rather than an immediate implementation step. While the modelling framework has been validated analytically and tested through target-oriented inverse use cases, its long-term value depends on the ability to embed model outputs into Harinera del Valle's operational systems. From an industrial perspective, deployment would require that formulation recommendations generated in the laboratory be transferred reliably and quickly to production plants across the country, and that resulting actions (dosage changes, production adjustments, and material consumption) are recorded in the company's ERP to ensure traceability, governance, and cost control.

A future deployment scenario therefore involves connecting three environments that are currently partially separated: (i) the laboratory environment where Mixolab and formulation decisions are generated, (ii) the operational management layer where plants execute production orders and manage process parameters, and (iii) the ERP environment where inventory movements, batch records, costing, and customer commitments are registered. In this integrated architecture, the model is positioned as a decision-support engine that translates laboratory evidence (flour characteristics, target dough profiles, and validation tests) into operational instructions that can be executed consistently in multiple plants, reducing reliance on informal communication or manual transcription.

A practical integration approach would follow a staged design. First, a standardized digital "formulation recommendation record" would be generated whenever the R&D team validates a target profile or runs an inverse-design recommendation. This record would include the flour batch identifier, the target Mixolab profile or client specification, the recommended additive dosing (ppm), and the expected Mixolab outcomes. Second, this record would be transmitted to the operations management system as a production instruction linked to a specific production order or client profile. Third, once production executes the dosing adjustments, the ERP would automatically register the material movements (enzyme, additive and overall materials consumption) and associate them with the batch produced. This closes the loop between technical intention (R&D recommendation), operational execution, and financial/traceability registration (ERP).

From a managerial perspective, this deployment idea supports a shift from local, experience-based formulation adjustments toward controlled and auditable decision-making. It paves early alignment between commercial agreements and technical feasibility by allowing R&D to generate a validated recommended profile before production begins, while at the same time finance and operations teams are working together to evaluate cost impacts using ERP-linked data. In addition, the accumulation of deployed recommendations and financial movements become a structured dataset that can be used for continuous learning, strengthening the long-term robustness of the framework.

5. Discussion

The results obtained in this study are broadly consistent with previous applications of machine-learning techniques in food processing, which report strong predictive performance for rheological and quality-related attributes when relevant compositional and process variables are available (Zhang et al., 2025; Pathare & Roskilly, 2016). Similar to earlier studies on cereal systems, the highest predictive accuracy was achieved for early and mid-stage dough properties, which are primarily driven by flour composition and formulation inputs that are explicitly measured and controlled (Delcour & Hoseney, 2010; Kotsiantis et al., 2007). The reduced predictability observed for slope γ, associated with enzymatic degradation during heating, aligns with prior findings highlighting the sensitivity of enzymatic phases to factors that are difficult to capture in routine industrial datasets, such as enzyme activity variability, microstructural starch differences, and subtle process-condition fluctuations (Goesaert et al., 2005; Van der Maarel et al., 2002).

Beyond predictive accuracy, this research extends the existing literature by demonstrating the practical feasibility of a stage-dependent and inverse modelling approach under real industrial conditions. While most prior studies focus on forward prediction or laboratory-scale optimization, the results presented here show that machine-learning models can be used to support target-oriented formulation decisions, generating realistic and bounded additive recommendations that align with industrial constraints. This shows how machine learning acts not only as a prediction tool but as a decision-support mechanism capable of reshaping complex workflows in specific contexts as a food industry in Latin America, with the potential to change commercial-technical interactions. The study contributes to an emerging research domain that links data-driven modelling with managerial decision-making in food manufacturing, highlighting how predictive analytics can reduce trial-and-error, support earlier feasibility assessments, and strengthen coordination between commercial, R&D, and production functions (Zhang et al., 2025; FAO, 2017).

6. Conclusion

The conclusion that follows reflects on how the results address the central research question and the specific objectives defined in the Introduction, and how the proposed analytical framework contributes to both formulation practices and managerial decision-making within Harinera del Valle.

The main research question of this study examined to what extent machine–learning, stage-dependent models can support formulation and managerial decision-making by reducing uncertainty and reliance on trial-and-error approaches in industrial flour production. The results demonstrate that the proposed modelling framework effectively contributes to this objective by enabling both accurate prediction of dough behaviour and actionable, target-oriented formulation support.

In relation to the first specific objective, which aimed to assess the extent to which machine-learning models can predict dough profiles based on flour characteristics and additive usage under industrial conditions, the findings confirm strong predictive performance across most Mixolab parameters. Stage-dependent Random Forest models achieved high accuracy, particularly for torque-related variables (C1, CS, C3, C4, and C5), with percentage MAE values consistently below 1%. These results indicate that the available formulation and compositional variables capture the main drivers of dough behaviour, supporting the feasibility of predictive modelling as a reliable tool in industrial contexts.

The second objective focused on identifying which stages of the dough rheology test are more reliably predictable and which are more sensitive to unobserved factors. The stage-wise evaluation revealed clear differences in predictability across the Mixolab sequence. Early and mid-stages (C1–C3) exhibited very high predictive accuracy, while later stages, showed significantly lower performance. This differentiation provides valuable insight into the process itself, highlighting where uncertainty is concentrated and where additional data may be required. From a managerial perspective, this enables interpretation knowing where the gaps stand and supports better informed decision-making.

The third objective examined whether a stage-dependent modelling framework can support targeted formulation by determining additive dosing prior to laboratory testing. The inverse, target-oriented analysis demonstrates that this is both feasible and operationally meaningful. Across multiple evaluated cases, the framework generated realistic adjustments in dosing, improving alignment with desired Mixolab profiles without producing impractical recommendations, as negative additives or high concentrations. This confirms that machine-learning models can be used not only for prediction but also as tools to formulate only with theoretical information, reducing reliance on iterative experimentation.

Taken together, these results provide a comprehensive answer to the central research question. The study shows that stage-dependent machine-learning models can significantly reduce uncertainty in formulation processes by enabling early prediction of dough behaviour, identifying stages with higher variability, and supporting the design of additive configurations aligned with target performance. With this, the framework changes the formulation process from a trial-and-error approach to a more structured system.

Beyond technical contributions, the findings highlight important managerial implications. The proposed framework supports a reconfiguration of the current operational and commercial workflow, enabling technical formulation decisions to be assessed prior to commercial commitment. This creates the possibility of alignment since step one between commercial, R&D, production, and financial functions, reducing risk of losing any potential client and improving coordination across the organization. In this sense, machine-learning models act as decision-support tools that enhance, rather than replace, expert judgment, providing a shared analytical basis for evaluating trade-offs between performance, cost, and feasibility.

Overall, this research demonstrates that integrating machine-learning models into industrial flour production can improve both formulation efficiency and decision-making processes. By aligning predictive analytics with the physical structure of dough transformation and the operational needs of the organization, the study contributes to a more data-driven, efficient, and responsible production system. In doing so, it provides a practical pathway for translating analytical capabilities into tangible industrial and managerial value.

7. LIMITATIONS AND FUTURE WORK

Despite the promising results obtained in this research along with the good outcome from the company, several limitations must be acknowledged. These limitations come from the nature of the data, the modelling approach, and the current stage of implementation within HV.

First, from a data and process perspective, the study is based on observations from a single company, Harinera del Valle, operating within a specific industrial and regional context. While this ensures strong practical relevance, it may limit possible generalizations of the findings to other production environments with different raw materials, processing technologies, or operational practices. In addition, the dataset does not explicitly capture properties of the data the model is ignoring by decision, which may influence dough behaviour at a more detailed level, the process itself is constrained by availability of data.

Future research could address these limitations by incorporating multi-site or multi-company datasets, enabling broader validation of the modelling framework across different industrial contexts. Additionally, the integration of more detailed process measurements, such as real-time sensor data or advanced compositional indicators, could improve model robustness and enhance the ability to capture currently unobserved sources of variability.

Second, the study presents model-related limitations inherent to the adopted analytical framework. The predictive models rely exclusively on observable inputs, meaning that latent or unmeasured factors affecting dough behaviour are not captured. This is particularly evident in stages such as enzymatic degradation, where lower predictive performance suggests the presence of hidden variables. Furthermore, the stage-dependent modelling approach introduces a degree of sequential error propagation, as predictions from earlier stages are used as inputs for subsequent models. While this structure reflects the physical logic of the Mixolab process, it also implies that inaccuracies in early-stage predictions may influence downstream results.

Future work could explore alternative modelling strategies to mitigate these limitations, including hybrid approaches that combine machine-learning models with domain-based constraints, or probabilistic methods capable of explicitly modelling uncertainty across stages. Additionally, expanding the modelling framework to include ensemble or hierarchical approaches may help reduce sensitivity to error propagation and improve predictive stability in later stages.

Finally, there are important implementation limitations. Although it has been validated and tested through numerous scenarios, it has not yet been fully deployed inside industrial operations. Its practical adoption requires integration with daily workflows such as laboratory procedures, production systems, and ERP platforms. This integration is not only technical but also organizational, as it implies changes in the overall process, this approach may require adjustments in company culture and internal dynamics between teams, also considering possible hierarchy.

Future work in this area should focus on industrial implementation, including pilot deployments within selected production lines, for example products which account for a small part of the revenue. Evaluating the impact of the framework with KPI´s such as formulation efficiency, cost control, and process variability will be essential to demonstrate its value. Additionally, studying user adoption and organizational adaptation will be critical to ensure that analytical tools effectively support the company decision loop in practice.

Overall, these limitations do not undermine the validity of the findings but rather define the scope within which the results should be interpreted. At the same time, they highlight a clear pathway for extending the research, both in terms of technical development and practical application, reinforcing the potential of data-driven approaches to transform formulation and decision-making processes in industrial flour production.

References

Aparicio, J.T., Aparicio, M., Costa, C.J. (2023). Design Science in Information Systems and Computing. In: Anwar, S., Ullah, A., Rocha, Á., Sousa, M.J. (eds) *Proceedings of International Conference on Information Technology and Applications. Lecture Notes in Networks and Systems*, vol 614. Springer, Singapore. https://doi.org/10.1007/978-981-19-9331-2_35

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chui, M., Manyika, J., & Miremadi, M. (2018). *What AI can and can't do (yet) for your business. McKinsey Quarterly. 1*(97-108), 1.

Codex Alimentarius Commission. (1995). *General standard for food additives (CODEX STAN 192-1995)*. FAO/WHO.

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In *Proceedings of the 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6). IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C. J. (2025). Introduction to Machine Learning. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.35c216ce

Courtin, C. M., & Delcour, J. A. (2002). Arabinoxylans and endoxylanases in wheat flour bread-making. *Journal of Cereal Science, 35*(3), 225–243. https://doi.org/10.1006/jcrs.2001.0433

Delcour, J. A., & Hoseney, R. C. (2010). *Principles of cereal science and technology* (3rd ed.). AACC International.

European Flour Millers & European Bakery Industry. (2020). *Sustainable wheat initiative: Roadmap to 2030.*

Food and Agriculture Organization of the United Nations. (2017). *SAVE FOOD: Global initiative on food loss and waste reduction*.

Food and Agriculture Organization of the United Nations. (2018). *Sustainable food systems: Concept and framework*.

Food and Agriculture Organization of the United Nations. (2019). *The state of food and agriculture 2019: Moving forward on food loss and waste reduction*.

Food and Agriculture Organization of the United Nations. (2023). *FAOSTAT food balances: Wheat utilisation*.

Food and Agriculture Organization of the United Nations, & World Health Organization. (2006). *General specifications and considerations for enzyme preparations used in food processing*. FAO/WHO Joint Expert Committee on Food Additives.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), https://doi.org/1189–1232. 10.1214/aos/1013203451

Goesaert, H., Slade, L., Levine, H., & Delcour, J. A. (2005). Amylases and bread firming—An integrated view. *Journal of Cereal Science, 41*(3), 305–321. https://doi.org/10.1016/j.jcs.2009.04.010

Harinera del Valle. (2025). *Corporate information and sustainability overview*. Internal company document.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering* (pp. 3–24). https://doi.org/10.5555/1566770.1566773

Kourtis, K., & MacGregor, J. F. (1995). Development of multivariate SPC charts using projection methods. *Chemometrics and Intelligent Laboratory Systems, 28*(1), 3–21. https://doi.org/10.1016/0169-7439(95)80036-9

Marin, D. (2026). *Stage-dependent machine-learning models for dough rheology* (v1.0) [Computer software]. GitHub. https://github.com/dmarinb99/Thesis-Project-MFW/blob/main/Final_Code.ipynb

McAfee, A., & Brynjolfsson, E. (2012). *Big data: The management revolution. Harvard Business Review*, 90(10), 60–68.

Pathare, P. B., Opara, U. L., & Al-Said, F. A.-J. (2013). Colour measurement and analysis in fresh and processed foods: A review. *Food and Bioprocess Technology, 6*(1), 36–60. https://doi.org/10.1007/s11947-012-0867-9

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science, 6*, e267. https://doi.org/10.7717/peerj-cs.267

Rosell, C. M., & Collar, C. (2008). Effect of various enzymes on dough rheology and bread quality. In *Recent research developments in food biotechnology: Enzymes as additives or processing aids* (pp. 165–183).

Shewry, P. R., & Halford, N. G. (2002). Cereal seed storage proteins: Structures, properties and role in grain utilization. *Journal of Experimental Botany, 53*(370), 947–958. https://doi.org/10.1093/jexbot/53.370.947

Tornberg, E. (2005). Quality and energy evaluation in meat cooking. In D. Sun (Ed.), *Emerging technologies for food processing* (pp. 323–344). Elsevier Academic Press. https://doi.org/10.1007/s12393-016-9143-5

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. https://sdgs.un.org/2030agenda

United Nations Environment Programme. (2021). *Making peace with nature*.

Van der Maarel, M. J. E. C., Van der Veen, B., Uitdehaag, J. C. M., Leemhuis, H., & Dijkhuizen, L. (2002). Properties and applications of starch-converting enzymes of the α-amylase family. *Journal of Biotechnology, 94*(2), 137–155. https://doi.org/10.1016/S0168-1656(01)00407-2

Zhang, L., Boom, R. M., & Ma, Y. (2025). Machine learning in automated food processing: A mini review. *Annual Review of Food Science and Technology, 16*, 25–37. https://doi.org/10.1146/annurev-food-111523-122039