UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Explainability of Machine Learning Models for Well-Being Prediction

Author: Leon Struman
Supervisor: Marco Saerens
Assistant: Flore Vancompernolle
Vromman
Academic year 2025-2026
Dissertation for the master of Business Engineering
Master subject and focus : Finance
Daytime schedule

Declaration regarding AI tool Usage

STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	**Confirmation: Yes**
The content of my/our master's thesis is my/our original input, even if I/we used generative AI to help prepare it.	☑
I/we master the theories, tools and methods that I/we use for the thesis.	☑
I/we verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☑
I/we acknowledge that I/we master the final output and that I am/we are able to explain it and answer questions about it.	☑

SIGNATURE

Sign your declaration:

	Author last name, first name(s)	**Date**	**Signature**
1	Leon Struman	26/05/2026	

Acknowledgements

I would first like to express my sincere gratitude to my thesis supervisor, Marco Saerens, for his support throughout the entire development of this work. His valuable advice, availability, and expertise were essential to the completion of this thesis.

I would also like to warmly thank Flore Vancompernolle Vromman, who guided me at the beginning of this project, particularly with the technical aspects. I am very grateful for the time she devoted to helping me and for her patience in answering my many questions.

Finally, I would like to thank everyone who contributed, directly or indirectly, to the completion of this work.

Contents

Introduction

Over the past decades, the concept of well-being has gradually taken a central place in many fields, ranging from public health and psychology to social sciences and, more recently, data science. Well-being is no longer considered only as the absence of illness, but as a multidimensional construct involving emotional, social, psychological, lifestyle, and contextual dimensions [17, 14]. Because of this complexity, its measurement and prediction remain challenging. Traditional approaches generally rely on validated questionnaires and expert interpretation, which are useful but can be time-consuming and difficult to deploy at a large scale [3, 17].

At the same time, recent advances in machine learning have opened new possibilities for estimating well-being and mental health-related outcomes from questionnaire data. Machine learning models can identify statistical patterns in large datasets and capture complex relationships between variables that may be difficult to model with classical approaches [13, 20, 3]. In this context, predicting a well-being score from a reduced set of lifestyle and contextual variables could support the development of more accessible tools for well-being assessment. However, this also raises an important methodological and practical challenge: the model must not only be accurate, but also understandable.

This issue is particularly important in domains related to health and well-being. A predicted well-being score is not a neutral technical output: it may influence how an individual understands their own situation, or how a professional interprets a given profile. Therefore, relying only on predictive performance is not sufficient. It is also necessary to understand which variables influence the prediction and whether the model's reasoning appears coherent, reliable, and interpretable [14, 11, 22]. This becomes especially problematic when complex machine learning models are used. Although black-box models, such as ensemble methods or boosting models, often achieve strong predictive performance, their internal decision mechanisms are difficult to interpret directly [2, 9, 11].

This tension between predictive performance and interpretability lies at the core of the present thesis. On the one hand, more complex models may improve the prediction of well-being by capturing non-linear relationships and interactions between lifestyle, social, health-related, and contextual variables. On the other hand, their lack of transparency may limit their practical relevance in a sensitive domain such as well-being. The challenge is therefore not only to determine whether black-box models can predict well-being effectively, but also whether their predictions can be explained in a meaningful way.

This thesis is situated within the broader Datagotchi Santé project, which aims to estimate individual well-being from a short questionnaire based on lifestyle and contextual variables [30]. The present work builds on this existing study and focuses on the comparison between interpretable linear models and more complex machine learning models, with a particular emphasis on explainability. A more detailed presentation of the dataset, the construction of the target variable, the preprocessing steps, and the initial modeling pipeline is provided in the following sections.

The primary objective of this thesis is therefore to examine whether non-linear and black-box machine learning models can improve the prediction of well-being compared to a linear reference model, while still producing explanations that can be interpreted. More specifically, this work evaluates several regression models and compares their predictive performance using standard error metrics. It then investigates the explainability of selected models using post-hoc XAI methods, with a particular focus on SHAP and LIME. These methods are widely used to explain predictions by quantifying the contribution of input features, either at the level of a single prediction or at the global model level [26, 19, 2].

From a management perspective, this question is also relevant because machine learning models are increasingly used to support decision-making in organizations. In such contexts, predictive performance alone is not sufficient: managers, decision-makers, and other stakeholders must also be able to understand, justify, and communicate the outputs of algorithmic systems. This is particularly important when AI-based tools are deployed in sensitive domains such as health and well-being, where predictions may influence user interpretation, service design, resource allocation, or organizational decisions. For a business engineering perspective, the issue is therefore not only technical, but also managerial: it concerns the conditions under which an organization can responsibly adopt a predictive model, balance accuracy with transparency, and ensure that data-driven tools remain trustworthy and usable in practice. In this

sense, the thesis contributes to the field of management by examining how explainability can support the responsible implementation of AI-based decision-support systems.

The research question guiding this thesis can therefore be formulated as follows:

> **Are black-box models explainable in the context of well-being prediction, and if so, how do they compare to the explainability of linear models?**

This question involves two complementary dimensions. The first concerns predictive performance: whether black-box models provide a meaningful improvement over simpler and more transparent models. The second concerns explainability: whether the explanations produced for these models are coherent, comparable across methods and models, and sufficiently reliable to support interpretation in a well-being context.

To address this question, the thesis follows an applied and comparative approach. First, several machine learning models are evaluated in order to identify whether more complex models improve the prediction of the well-being score. Second, the best-performing and most relevant models are analyzed using explainability methods. The analysis compares global and local explanations, examines the similarity between explanation patterns, and evaluates selected aspects of explanation reliability, namely faithfulness and robustness. These dimensions are relevant because recent XAI literature emphasizes that explanations should not only be intuitive, but also faithful to the model's behavior and stable under reasonable perturbations [1, 10, 4].

Overall, this thesis contributes to the broader discussion on the use of explainable artificial intelligence in well-being prediction. Its objective is not only to identify the model with the lowest prediction error, but also to assess whether the use of black-box models is justified when interpretability is considered. In this sense, the work aims to provide a critical evaluation of the balance between accuracy and explainability in a sensitive, real-world prediction context.

Chapter 1

Literature review

1.1 Well-being modeling and XAI

Psychological well-being is a complex and multidimensional concept, encompassing emotional, social, behavioral, and cognitive dimensions [17]. Unlike physiological indicators that can be measured objectively, well-being relies primarily on subjective perceptions, such as emotional state, perceived life satisfaction, the quality of social relationships, and the ability to cope with life events [17]. This inherently subjective nature makes its quantification particularly challenging and explains why, historically, its assessment has relied on standardized questionnaires and the expertise of mental health professionals [17, 14].

Many studies in mental health rely on validated psychometric instruments, such as anxiety, depression, or stress scales, to capture different facets of well-being. For example, Byeon [3] uses the Zung Self-Rating Anxiety Scale (SAS), a clinically well-validated questionnaire, to investigate factors associated with anxiety disorders among older adults living alone. This type of instrument allows complex psychological dimensions to be measured through a structured set of questions, but at the cost of relatively lengthy questionnaires and interpretations that often remain dependent on clinical judgment [3, 17, 14].

However, these traditional approaches present several limitations. First, they require a substantial time investment from both participants and professionals, which complicates their deployment at scale [3, 13]. Second, the interpretation of results often relies on human analysis, which can introduce inter-rater variability [14]. Finally, the relationships between the different dimensions of well-being are rarely linear: social, economic, and individual factors interact in complex ways, making them difficult to model using classical statistical methods [13, 3].

The work of Jha et al. [13] clearly illustrates this complexity in the context of the COVID-19 pandemic. Using a large dataset derived from surveys conducted among nearly 18,000 adults in the United States, the authors show that mental well-being is simultaneously influenced by demographic factors, working conditions, social dynamics, medical history, and contextual constraints. Their study highlights that these factors cannot be analyzed in isolation without losing a significant portion of the information, thereby underscoring the limitations of univariate or purely linear analyses for understanding well-being.

This difficulty is also evident in studies focusing on specific populations, such as children and adolescents. Ntakolia et al. [20] demonstrate that the impact of the pandemic on the mood of children and adolescents in Greece results from a heterogeneous set of variables spanning multiple categories, including family life, daily habits, health status, social context, and psychiatric history. The authors emphasize that no single variable, when considered in isolation, is sufficient to explain the observed variations in well-being, reinforcing the idea that well-being should be understood as the outcome of complex interactions among multiple factors.

In light of these observations, the use of machine learning models appears to be a relevant alternative to traditional approaches. Machine learning methods are indeed capable of capturing non-linear relationships and complex interactions between variables while leveraging high-dimensional datasets. Recent studies show that these approaches can significantly improve the predictive performance of well-being or mental health states compared to classical statistical models [3, 13, 20]. However, these performance gains

are generally accompanied by a loss of transparency, as the most accurate models are often described as black boxes.

This lack of transparency is particularly problematic in applications related to well-being and mental health. In such contexts, predictions are not neutral technical outputs: they may contribute to the interpretation of an individual's psychological state, support professional judgment, or guide decisions related to care and prevention. Therefore, evaluating a model only through predictive performance is insufficient. It is also necessary to understand which factors influence the prediction and whether the reasoning followed by the model appears coherent from a clinical, psychological, and ethical perspective.

This is where Explainable Artificial Intelligence (XAI) becomes central. XAI makes it possible to go beyond a purely quantitative evaluation of model performance by providing insight into the factors underlying predictions. In mental health and well-being contexts, explainability can support the reliability and acceptability of machine learning models by helping clinicians, experts, and users understand how predictions are produced [14, 11]. Several studies also emphasize that explanations can strengthen trust in AI systems, especially when they reveal decision mechanisms that are consistent with existing medical or psychological knowledge [12, 11].

Moreover, XAI is not limited to improving user trust. It also constitutes an important tool for the critical validation of machine learning models. By making model reasoning more explicit, explainability can help identify incoherent dependencies, irrelevant variables, potential biases, or artifacts in the data [14, 22]. This is particularly important in sensitive domains such as well-being, where an apparently accurate model may still rely on questionable patterns or produce explanations that are difficult to justify from a domain-specific perspective.

Thus, the modeling of well-being currently lies at the intersection of two sometimes competing requirements: on the one hand, the need for sufficiently expressive models to capture the complexity of the phenomenon, and on the other hand, the obligation to provide understandable and interpretable results, particularly in sensitive health-related contexts. This tension represents a central challenge for recent work in artificial intelligence applied to well-being and justifies the growing interest in approaches that seek to combine predictive performance with explainability.

1.2 Machine learning models for prediction

1.2.1 Nature of the data and problem formulation

Data used in studies on well-being and mental health primarily originate from self-reported questionnaires or questionnaires administered by professionals [3, 20, 13]. These instruments cover heterogeneous dimensions, such as sociodemographic characteristics, lifestyle habits, psychological symptoms, social context, and medical history [13, 20]. Ntakolia et al. [20] emphasize that predicting well-being status relies on a combination of variables drawn from multiple categories, and that no single variable, when considered in isolation, is sufficient to explain the observed variations.

In this context, the problem can be formulated either as a regression task, when the objective is to predict a continuous well-being score, or as a classification task, when the goal is to predict severity categories or the presence of a disorder. The methodological constraints remain similar in both cases: strong correlations among certain variables, non-linear interactions, and complex dependencies between psychosocial factors. Jha et al. [13] highlight these dependencies using Bayesian networks, showing that factors influencing mental health cannot be analyzed independently without a loss of information.

1.2.2 Data preprocessing: general framework

Questionnaire data generally require preprocessing steps before they can be used in machine learning models. These steps are particularly important when the dataset contains heterogeneous variables, missing values, categorical answers, and features measured on different scales. In the context of well-being and mental health prediction, preprocessing commonly includes the transformation of questionnaire responses into numerical variables, missing value imputation, feature scaling, and variable selection [15].

In the Datagotchi Santé project, these preprocessing steps were carried out upstream and are described in detail by Vancompernolle Vromman et al. [30]. The initial questionnaire contained 280 items. To ensure that the predictive model focused on lifestyle and contextual indicators rather than on direct measures of well-being, the items used to compute the target variable were excluded. After this filtering step, 74 questionnaire items were retained for feature construction. Because some questionnaire items were categorical or multicategorical, the preprocessing stage expanded these 74 items into a larger feature library of 214 engineered features.

The preprocessing pipeline was integrated within the cross-validation procedure in order to avoid data leakage. Within each training fold, missing values were imputed, features were scaled, feature selection was applied, and the model was then trained. This means that preprocessing decisions were learned only from the training data and evaluated on held-out test data. This is important because applying preprocessing before the train-test split could lead to overly optimistic performance estimates.

A feature selection procedure was also applied to reduce the number of predictors and make the final questionnaire more practical for deployment. Starting from the 214 engineered features, the procedure selected a smaller subset of high-value predictors. In the final model presented by Vancompernolle Vromman et al. [30], 20 questionnaire items were retained. Due to multicategory variables, these 20 items corresponded to 36 engineered features in the final modeling dataset.

Missing values were handled through imputation. In the final refit of the ridge regression model reported in the original project, the selected strategy was a k-nearest neighbors imputer with five neighbors. Feature scaling was performed using a MinMaxScaler. These preprocessing choices are not the main focus of the present thesis, but they are essential because they make the questionnaire data usable for different machine learning models and ensure that model comparisons are based on a consistent input representation.

1.2.3 Linear models

Linear models often constitute a natural starting point for well-being prediction [2, 11]. In regression settings, these models assume that the target variable can be expressed as a linear combination of the input variables. Regularization techniques such as Ridge or Lasso regression are commonly employed to mitigate overfitting and handle multicollinearity among features [8].

The main advantage of these models lies in their intrinsic interpretability: each coefficient is directly associated with the impact of a given variable on the prediction [2, 8, 14]. However, several studies show that this simplicity comes at the cost of limited expressive power. Kaur et al. [15] observe that linear models achieve lower performance than non-linear models when relationships between variables are complex and non-additive. Similarly, Ntakolia et al. [20] report weaker performance for logistic regression compared to more advanced ensemble models.

1.2.4 Non-linear and black-box models

Although linear models provide an interpretable reference point, they may be too limited to capture the complex and non-linear relationships involved in well-being prediction [15, 20]. Psychological, social, lifestyle, and contextual variables can interact in ways that are not purely additive [13, 3]. For this reason, non-linear machine learning models such as gradient boosting methods, support vector regressors, and neural networks are often used to improve predictive performance [5, 15, 3]. However, their higher flexibility generally comes at the cost of reduced transparency, which is why they are commonly described as black-box models [2, 9, 11]. This subsection therefore presents some of the most widely used black-box models discussed in the literature, with a focus on their underlying mechanisms and relevance for well-being prediction.

Ensemble learning and stacking approaches

Before presenting specific black-box models, it is useful to introduce the broader family of ensemble methods. Ensemble learning refers to approaches that combine several individual predictors in order to produce a single final prediction [3, 5, 16]. The underlying idea is that different models, or different

versions of the same model, may capture complementary patterns in the data. By combining them, ensemble methods can often improve predictive performance and robustness compared to a single model [3].

Several black-box models used in machine learning belong to this family. For example, Random Forest combines multiple decision trees trained in parallel, while gradient boosting methods such as XGBoost, LightGBM, and CatBoost build an ensemble of trees sequentially, with each new tree correcting part of the error made by the previous ensemble [5, 16, 24]. These approaches are particularly relevant for well-being prediction because psychological, social, lifestyle, and contextual variables may interact in complex and non-linear ways [13, 20, 3].

Stacking, also called stacked generalization, is another ensemble approach. Unlike bagging or boosting, stacking combines different types of models by using their predictions as inputs for a second-level model. Ting and Witten [28] explain that the first models, often called level-0 models or base models, are trained on the original input variables. Each of these models produces a prediction. These predictions are then collected and used to form a new dataset, which is used to train another model, called the level-1 generalizer or meta-model. The role of this meta-model is to learn how to combine the predictions of the base models in order to produce the final prediction [28].

In a regression context, the logic of stacking can be summarized as follows:

$$\hat{y} = g(f_1(x), f_2(x), \ldots, f_M(x)) \tag{1.1}$$

where $f_1, f_2, \ldots, f_M$ represent the base models, x denotes the input variables, and g is the meta-model that learns how to combine their predictions. For example, the base models may include a gradient boosting model, a support vector regressor, or a neural network, while the meta-model may be a linear regression or another machine learning model [5, 27, 23]. Instead of simply averaging the predictions, stacking allows the model to learn which base learners should be trusted more in different situations [28].

A central methodological point in stacking is that the meta-model should not be trained on predictions obtained from base models evaluated on the same data used to fit them. If this were the case, the meta-model could learn from over-optimistic predictions and overfit the training data. For this reason, stacking is usually implemented using cross-validation: the training data are split into folds, base models are trained on part of the data, and their predictions are generated on held-out folds. These out-of-sample predictions are then used to train the meta-model [28].

Stacking is relevant in mental health and well-being prediction because different algorithms may capture different aspects of complex psychological and lifestyle data [3, 13, 20]. For instance, Byeon [3] uses a stacking ensemble for anxiety disorder prediction, combining SVM, Random Forest, LightGBM, and AdaBoost as base models, with XGBoost as the meta-model. Their results show that the stacking ensemble achieved the best predictive performance among the tested approaches, illustrating the potential value of combining complementary models in mental health applications [3].

However, stacking also increases model complexity. Since the final prediction depends both on several base models and on an additional meta-model, the resulting system is generally less transparent than a single model [28, 9, 2]. This makes stacking particularly difficult to interpret directly and reinforces its classification as a black-box approach. Consequently, when stacking is used in well-being prediction, explainability methods are needed to understand not only which variables influence the final prediction, but also how the different base models contribute to it [2, 9, 14].

Gradient boosting models

Gradient boosting is a widely used ensemble learning approach based on the iterative combination of weak predictors, generally decision trees, into a stronger predictive model [5, 16, 24]. Among black-box models, gradient boosting refers to a family of methods that includes several widely used algorithms such as XGBoost, LightGBM, and CatBoost. The principle of this family is not to fit a single predictive model, but to combine several weak learners, most often decision trees, into a stronger model. These trees are trained sequentially: each new tree is added to correct the errors made by the current ensemble. In regression settings, the model starts from an initial prediction and progressively adds trees that reduce the remaining prediction error. More formally, the final prediction can be expressed as an additive model,

where the output is obtained by summing the predictions produced by all trees in the ensemble [5, 24]. This ability to combine multiple decision rules makes gradient boosting particularly suitable for capturing non-linear relationships and interactions between variables, which is relevant in well-being prediction, where lifestyle, psychological, social, and contextual factors may interact in complex ways.

The training process of gradient boosting decision trees relies on the minimization of a loss function, where each new tree is fitted using information derived from the gradient of this loss [16, 24]. At each iteration, the algorithm uses the gradient of this loss function to determine how the current model should be improved. A new decision tree is then fitted to approximate this correction. In other words, each tree does not learn the target variable from scratch, but learns how to improve the current prediction. This sequential correction mechanism explains why gradient boosting models can achieve high predictive performance. However, because the final prediction depends on the combination of many trees, splits, and interactions, these models are generally difficult to interpret directly [9, 2]. The following subparagraphs present three major models from this family: XGBoost, LightGBM, and CatBoost.

XGBoost Chen and Guestrin [5] introduce XGBoost as a scalable implementation of gradient tree boosting. In their formulation, the model is defined as an ensemble of regression trees, where each tree maps an observation to a leaf and each leaf contains a numerical score. The final prediction is obtained by summing the scores assigned by all trees in the ensemble:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\hat{y}_i$ denotes the predicted value for observation i, K is the number of trees, and each f_k represents one regression tree. Therefore, in a regression task, XGBoost predicts a continuous outcome by accumulating the contributions of multiple regression trees.

A key contribution of XGBoost is the use of a regularized learning objective. This objective contains two components: a loss term, which measures the difference between predicted and observed values, and a regularization term, which penalizes the complexity of the trees:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

with

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

In this formulation, $l(\hat{y}_i, y_i)$ represents the prediction error, T is the number of leaves in a tree, and w denotes the vector of leaf weights. The parameters γ and λ control the strength of the regularization. Intuitively, this term discourages overly complex trees by penalizing both the number of leaves and the magnitude of the leaf scores. Its purpose is to reduce the risk of overfitting, which is particularly important in datasets with many variables or limited sample sizes, where a flexible model may otherwise learn noise rather than stable patterns [5].

XGBoost also introduces several algorithmic optimizations to improve computational efficiency and scalability. Chen and Guestrin [5] describe, among others, a sparsity-aware algorithm for handling missing or sparse data, a weighted quantile sketch for approximate split finding, and system-level optimizations such as cache-aware computation and out-of-core learning. These features make XGBoost applicable to large datasets while maintaining strong predictive performance.

LightGBM Ke et al. [16] introduce LightGBM as a gradient boosting decision tree algorithm designed to improve the efficiency and scalability of conventional GBDT methods. While LightGBM follows the same general boosting logic as other gradient boosting models, its main contribution lies in the way it accelerates tree construction. The authors explain that traditional gradient boosting implementations become computationally expensive when the number of observations and features is large, because the algorithm must scan many possible split points for each feature. LightGBM addresses this issue mainly through two techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

GOSS is based on the idea that observations with large gradients are more informative for learning. A large gradient indicates that the model currently predicts the observation poorly, meaning that this observation provides useful information for improving the model. Instead of using all observations to estimate the best splits, GOSS keeps the observations with large gradients and samples only a subset of observations with small gradients. This reduces the amount of data used during training while preserving much of the information needed to improve the model [16].

EFB aims to reduce the number of features. In sparse datasets, many features are mutually exclusive, meaning that they rarely take non-zero values at the same time. This situation is common after one-hot encoding categorical variables. LightGBM bundles such mutually exclusive features together, thereby reducing the effective dimensionality of the data without substantially affecting split quality. Ke et al. [16] show that the combination of GOSS and EFB allows LightGBM to train much faster than conventional GBDT implementations while maintaining almost the same predictive accuracy.

Another important characteristic of LightGBM is its leaf-wise tree growth strategy. Unlike level-wise algorithms, which grow all leaves at the same depth, LightGBM selects the leaf that provides the largest reduction in loss and splits it first. This can lead to stronger loss reduction and better predictive performance, but it may also increase the risk of overfitting if the tree depth or number of leaves is not properly controlled [16]. Byeon [3] also highlights this leaf-wise strategy in the context of anxiety disorder prediction and notes that LightGBM can reduce loss more efficiently than level-wise boosting algorithms.

CatBoost Prokhorenkova et al. [24] introduce CatBoost as a gradient boosting algorithm designed to handle categorical variables more effectively and to reduce prediction bias. This is particularly relevant for questionnaire-based datasets, where many variables are categorical or ordinal. A common approach for categorical variables is target encoding, where each category is replaced by a statistic computed from the target values of observations belonging to that category. However, if this statistic is computed using the full training set, the target value of an observation may indirectly be used to encode its own features, creating target leakage.

To address this issue, CatBoost uses ordered target statistics. The idea is to compute the encoding of a categorical value using only previous observations in a random permutation of the dataset. A simplified formulation is:

$$\hat{x}_k^i = \frac{\sum_{j=1}^{k-1} \mathbf{1}\{x_j^i = x_k^i\} y_j + aP}{\sum_{j=1}^{k-1} \mathbf{1}\{x_j^i = x_k^i\} + a}$$

where $\hat{x}_k^i$ is the encoded value of categorical feature i for observation k, y_j is the target value of a previous observation with the same category, P is a prior value, and a controls the strength of this prior. The important point is that the encoding of observation k is based only on observations that appear before it in the permutation, which prevents the model from using the current observation's own target value during training [24].

CatBoost also introduces ordered boosting, a permutation-based version of gradient boosting that aims to reduce prediction shift. In standard boosting, the gradients used to train a new tree may be computed from a model that has already used the same observations during training. CatBoost reduces this issue by ensuring that the prediction used to compute the gradient for an observation is obtained from a model trained only on previous observations in the permutation [24]. As a result, CatBoost is particularly useful when datasets contain categorical variables.

Support Vector Regressor

Support Vector Regression (SVR) is the regression extension of Support Vector Machines. Smola and Schölkopf [27] explain that the objective of SVR is to find a function that predicts the target values with a maximum tolerated error of ε, while remaining as flat as possible. In other words, small prediction errors are ignored as long as they remain inside an ε-margin around the predicted function. Only observations located outside this margin contribute to the loss. This principle is known as the ε-insensitive loss function.

In the linear case, SVR estimates a function of the form:

$$f(x) = \langle w, x \rangle + b$$

where w represents the vector of coefficients, x the input features, and b the intercept. The notion of flatness is obtained by minimizing the norm of w, which encourages a simpler function. However, because it is often impossible to fit all observations within the ε-margin, slack variables are introduced to allow some deviations. The optimization problem can therefore be written as:

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to prediction errors being smaller than ε plus the slack variables. The parameter C controls the trade-off between model flatness and tolerance for errors larger than ε. A high value of C penalizes large errors more strongly, while a lower value allows a smoother but potentially less accurate function [27].

SVR can also model non-linear relationships through the kernel trick. Instead of fitting a linear function in the original feature space, the model implicitly maps the data into a higher-dimensional space where non-linear patterns can be captured. This makes SVR relevant for well-being prediction, where the relationship between questionnaire variables and the well-being score may not be purely linear. However, when non-linear kernels are used, the resulting model becomes harder to interpret directly, since predictions depend on support vectors and kernel similarities rather than explicit coefficients. For this reason, SVR is generally considered a black-box model in non-linear applications.

Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a type of artificial neural network composed of an input layer, one or more hidden layers, and an output layer. Popescu et al. [23] describe the MLP as a feed-forward neural network, meaning that information moves in one direction, from the input variables to the output prediction, without feedback loops. In a regression task, the output layer produces a continuous predicted value, which makes MLPs applicable to the prediction of well-being scores.

The basic unit of an MLP is the neuron. Each neuron receives several input values, computes a weighted sum of these inputs, adds a bias term, and then applies an activation function. This can be written as:

$$z = \sum_{j=1}^{p} w_j x_j + b$$

$$a = \sigma(z)$$

where x_j represents the input features, w_j the associated weights, b the bias term, and $\sigma(\cdot)$ the activation function. The use of non-linear activation functions is essential, because Popescu et al. [23] explain that a network composed only of linear activation functions would not provide more computational power than a single linear model. Non-linear activation functions allow MLPs to approximate complex relationships between inputs and outputs.

MLPs are trained by adjusting their weights in order to reduce the difference between predicted and observed values. The most common training procedure is the backpropagation algorithm, which propagates the prediction error backward through the network and updates the weights using a gradient-based optimization method [23]. This learning process allows the model to capture non-linear effects and interactions between variables, which can be useful in well-being prediction when psychological, social, and lifestyle factors interact in complex ways.

However, MLPs also present important limitations. Ramchoun et al. [25] emphasize that the choice of architecture, including the number of hidden layers, neurons, and connections, has a strong impact on model performance. Too few connections may prevent the network from solving the problem, while too many connections may lead to overfitting. In addition, MLPs are generally less transparent than linear models because their predictions result from many weights, hidden units, and non-linear transformations. For this reason, they are often considered black-box models and may require post-hoc explainability methods to interpret their predictions.

1.3 Performance of models

The machine learning models presented above are useful because they can capture different types of relationships in the data and may improve prediction quality. However, the literature does not evaluate these models only in theoretical terms; their relevance is generally assessed through empirical performance metrics. For example, studies in mental health prediction commonly compare models using indicators such as accuracy, precision, recall, F1-score, or error-based measures, depending on whether the task is formulated as a classification or regression problem [3, 15, 20]. These metrics make it possible to compare competing models and identify which approach provides the most reliable predictions for a given dataset.

In the context of this thesis, the prediction task is formulated as a regression problem, since the objective is to estimate a continuous well-being score. The models that have already been tested in the project are therefore evaluated using two regression metrics: the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). These two indicators provide complementary information about prediction error and will be explained in the following subsections.

1.3.1 Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2)

Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) are commonly used evaluation metrics for regression models [6]. Since the objective of this thesis is to predict a continuous well-being score, these metrics make it possible to quantify how close the predicted values are to the observed values and to compare the predictive performance of the different models. In machine learning studies related to mental health and well-being, predictive models are commonly evaluated through empirical performance metrics adapted to the nature of the prediction task, such as classification metrics for categorical outcomes or regression metrics for continuous outcomes [3, 13, 20].

The Mean Squared Error (MSE) measures the average squared difference between the predicted values and the observed values [6]. For a dataset containing n observations, it can be written as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{1.2}$$

where y_i represents the observed value for observation i, and $\hat{y}_i$ represents the corresponding predicted value. Lower MSE values indicate better predictive performance, while a value of 0 corresponds to perfect predictions [6]. Because the errors are squared before being averaged, MSE gives more weight to large prediction errors. This makes it useful for identifying models that avoid large deviations between predicted and observed well-being scores [6].

The Mean Absolute Error (MAE) measures the average absolute difference between the predicted values and the observed values [6]. It can be written as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1.3}$$

Like MSE, lower MAE values indicate better predictive performance, and a value of 0 corresponds to perfect predictions [6]. However, MAE is easier to interpret because it is expressed in the same unit as the target variable [6]. In the context of this thesis, the target variable is a well-being score on a 0–100 scale. Therefore, an MAE of 14 means that, on average, the model prediction differs from the observed well-being score by approximately 14 points.

In addition to these error-based metrics, the coefficient of determination, denoted as R^2, can also be used to evaluate regression models [6]. Contrary to MSE and MAE, which measure prediction error, R^2 indicates the proportion of variance in the target variable that is explained by the model [6]. It can be written as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.4}$$

where $\bar{y}$ represents the mean of the observed values. An R^2 value close to 1 indicates that the model explains a large proportion of the variance in the target variable, while a value close to 0 indicates that the model does not perform much better than a simple baseline model that always predicts the mean of the observed target values [6]. Negative values can also occur when the model performs worse than this mean-based baseline [6].

These three metrics provide complementary perspectives on model performance. MSE is sensitive to large errors and therefore penalizes models that occasionally make strong prediction mistakes. MAE provides a more direct interpretation of the average prediction error on the well-being score scale. R^2, by contrast, gives an indication of how much of the variability in well-being scores is captured by the model [6]. This complementarity is important because model evaluation should not rely on a single metric only, especially in health-related applications where predictive performance must be interpreted carefully [22].

In this thesis, MAE and MSE are retained as the primary metrics for comparing predictive performance, because they directly quantify the prediction error. MAE is particularly useful because it can be interpreted directly on the 0–100 well-being score scale, while MSE highlights models that produce larger errors. The R^2 score is used as a complementary indicator, as it provides additional information about the proportion of variance explained by each model [6].

1.3.2 Statistical tests for model comparison

In addition to reporting performance metrics such as MSE and MAE, the literature emphasizes that model evaluation should not rely only on raw metric values. Chicco, Warrens, and Jurman [6] note that regression metrics such as MSE and MAE are widely used, but that their values are meaningful mainly in relation to the scale of the target variable and when compared across models. In the context of health-related machine learning, Pendyala and Kim [22] further argue that conventional evaluation metrics may indicate model performance but do not necessarily guarantee the reliability of the model outcomes. This motivates the use of complementary analyses when comparing predictive models.

Statistical testing can therefore be used to assess whether observed differences in model performance are sufficiently consistent to support a meaningful comparison. In the machine learning literature, performance is often estimated through cross-validation: for example, Byeon [3] evaluate several mental-health prediction models using cross-validation and compare their performance using evaluation metrics. In a similar logic, the present thesis evaluates each model using cross-validation folds. Since MSE and MAE are obtained separately for each fold, statistical tests can be applied to these fold-level errors to examine whether the difference between two models is systematic rather than only due to variation across folds.

In the following subsections, we briefly introduce the statistical tests that are most commonly considered for comparing models evaluated on the same data splits.

Nemenyi post-hoc test.

Trawinski et al. [29] present the Nemenyi test as a non-parametric post-hoc procedure used for multiple comparisons between machine learning regression algorithms. In their framework, a global test such as the Friedman test or the Iman–Davenport test is first applied to determine whether significant differences exist among the algorithms. If this global null hypothesis is rejected, post-hoc procedures can then be used to identify which specific pairs of algorithms differ significantly. The Nemenyi test is one of these post-hoc procedures and is designed for $N \times N$ comparisons, meaning that all possible pairs of algorithms are compared [29].

The Nemenyi test is based on the average ranks obtained by the algorithms across the different evaluation blocks, such as datasets or cross-validation folds. In each block, the algorithms are ranked according to their predictive performance, and these ranks are then averaged over all blocks. Two algorithms are considered significantly different if the difference between their average ranks is larger than a threshold called the *critical difference*. The critical difference therefore represents the minimum distance required between two average ranks for their performance difference to be considered statistically significant at a chosen significance level.

The critical difference is computed as follows:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{1.5}$$

where k is the number of algorithms compared, N is the number of evaluation blocks, and q_α is a critical value derived from the Studentized range distribution for the chosen significance level. A larger number of algorithms or a smaller number of evaluation blocks increases the critical difference, making the test more conservative. In practical terms, if the absolute difference between the average ranks of two models is greater than the critical difference, the Nemenyi test concludes that the two models differ significantly. If the difference is smaller, no statistically significant difference is detected.

Trawinski et al. [29] explain that the Nemenyi test adjusts the significance level in a single step according to the total number of pairwise comparisons. This makes the method simple to apply, but also relatively conservative and less powerful than alternatives such as Shaffer or Bergmann–Hommel. In the context of this thesis, the Nemenyi test can therefore be understood as a tool for comparing several models simultaneously based on their average ranks. However, its results should be interpreted cautiously because the present analysis is based on cross-validation folds from a single dataset rather than on multiple independent datasets.

Wilcoxon signed-rank test

Trawinski et al. [29] present the Wilcoxon signed-rank test as a nonparametric test commonly used for pairwise comparisons between two machine learning algorithms. Unlike multiple comparison procedures such as the Friedman or Nemenyi tests, the Wilcoxon test is designed to compare only two models at a time. It evaluates whether the performance differences between two algorithms, measured on the same datasets or evaluation splits, are systematically different from zero [29]. In this sense, it is a paired test: each performance value of one model is directly compared with the corresponding performance value of the other model on the same evaluation unit.

Trawinski et al. [29] explain that the Wilcoxon test ranks the absolute differences between the two algorithms' performances, while taking into account the sign of each difference. The test then compares the ranks associated with positive and negative differences. If one model consistently obtains lower error values than the other, the ranks will tend to favour that model, suggesting a significant difference in performance. Compared with the paired t-test, the Wilcoxon test does not require the assumption of normally distributed differences, which makes it useful when this assumption is difficult to justify [29]. However, Trawinski et al. [29] also emphasize that the Wilcoxon test is appropriate for pairwise comparisons only. When several models are compared simultaneously, applying multiple Wilcoxon tests can increase the risk of false positive conclusions because the family-wise error rate is no longer controlled.

Other statistical tests for model comparison

Beyond the Wilcoxon and Nemenyi tests, several other statistical procedures are commonly discussed in the machine learning literature for comparing model performance. Trawinski et al. [29] distinguish between pairwise tests, which compare two algorithms, and multiple-comparison tests, which are designed to compare several algorithms simultaneously. For global comparisons involving more than two models, they present the Friedman test as a nonparametric alternative to analysis of variance. This test ranks the algorithms on each dataset or evaluation unit and then assesses whether the average ranks differ significantly across models. However, the Friedman test only indicates whether at least one model differs from the others; it does not identify which specific pairs are significantly different [29]. Trawinski et al. [29] also discuss the Iman–Davenport test, which is a more powerful modification of the Friedman test and is used for the same purpose: detecting global differences among several algorithms.

For comparisons between one reference model and several alternatives, Trawinski et al. [29] mention post-hoc correction procedures such as Bonferroni–Dunn, Holm, Hochberg, Hommel, Finner, and Li. These methods control the risk of false positive conclusions when multiple hypotheses are tested. For all pairwise comparisons between models, they discuss procedures such as Nemenyi, Shaffer, and Bergmann–Hommel. Among these, Nemenyi is simple but relatively conservative, while Shaffer and Bergmann–Hommel are more powerful but also more complex [29].

1.4 Interpretability and explainability of models

The notions of interpretability and explainability are central to the evaluation and deployment of machine learning models, particularly when they are used in sensitive domains such as health and well-being. The literature highlights that there is a clear conceptual distinction between these two notions, although they are often used interchangeably.

Interpretability refers to an intrinsic property of a model, corresponding to its ability to be directly understood by a human without relying on external explanatory mechanisms. An interpretable model allows the explicit identification of the relationship between input variables and the resulting prediction, for instance through coefficients or explicit rules, as is the case for linear models or shallow decision trees [2, 9].

In contrast, explainability is considered an extrinsic or active property, referring to the set of methods and procedures aimed at explaining the behavior of a model that is not interpretable by design. These so-called post-hoc methods seek to provide a human-understandable approximation of the internal reasoning of a complex model, without modifying its architecture or internal functioning [2, 9].

These black-box models, such as random forests, gradient boosting methods, or deep neural networks, are characterized by high expressive power and superior predictive performance, but at the cost of significant opacity in their decision-making processes. This opacity prevents direct interpretation of predictions and raises issues related to trust, clinical validation, and accountability, particularly in applications related to mental health and well-being [9, 14].

The literature thus converges on the idea that there is a structural trade-off between performance and transparency: intrinsically interpretable models offer direct but limited understanding of complex phenomena, whereas black-box models achieve higher performance at the expense of interpretability. Explainable Artificial Intelligence (XAI) emerges precisely to address this trade-off by proposing mechanisms that make the predictions of complex models understandable without sacrificing predictive efficiency [2, 9].

1.5 XAI methods

Explainable Artificial Intelligence (XAI) encompasses a set of methods aimed at making the predictions produced by complex machine learning models—often referred to as black boxes—understandable. As these models reach high levels of performance, particularly in critical domains such as health and well-being, predictive accuracy alone is no longer sufficient to ensure reliability, acceptability, and responsible use. XAI therefore serves as a central lever to enhance transparency, detect undesirable behaviors, and support human decision-making [2, 11].

1.5.1 Selection of XAI methods according to context and audience

Before selecting specific XAI methods, it is important to note that explainability is not a single, universal objective. The relevance of an explanation depends on the context in which it is used, the type of model being explained, the nature of the data, and the audience receiving the explanation [2, 4]. Barredo Arrieta et al. [2] emphasize that explainability should be understood in relation to a given audience, because the same explanation may not be equally useful for a data scientist, a domain expert, a regulator, or an end user. Similarly, Doshi-Velez and Kim [7] argue that interpretability needs vary across applications and that explanation methods should therefore be evaluated according to the task and the users for whom they are intended. In this sense, an XAI method should not only be selected because it is technically available, but because it produces a form of explanation that is adapted to the practical objective of the analysis [7, 2].

A first important distinction concerns the scope of the explanation. Global methods aim to explain the overall behavior of a model across the dataset. They are useful when the objective is to understand the general logic of the model, identify dominant variables, audit the model, or detect potential biases [2, 4]. In contrast, local methods focus on a single prediction and seek to explain why the model produced a specific output for a given individual [2, 4]. These explanations are particularly relevant in

decision-support contexts, where understanding the reasoning behind an individual prediction may be more important than only knowing the model's average behavior [7, 2]. In the context of this thesis, both perspectives are useful: global explanations help identify the main variables influencing well-being prediction at the population level, while local explanations make it possible to analyze how these variables contribute to the predicted score of a specific individual.

A second structuring distinction concerns the level of model dependency. Model-specific methods rely on the internal structure of a particular model family, such as tree structures, coefficients, gradients, or neural network weights [4]. These methods can provide explanations that are closely connected to the model mechanism, but their use is limited to compatible model architectures [4]. By contrast, model-agnostic methods treat the predictive model as a black box and only require access to its inputs and outputs [4, 1]. This makes them more flexible, especially when several types of models must be compared [1]. This distinction is particularly important in this thesis, since the analysis compares different families of models, including linear models, boosting models, and stacking approaches.

The type of data and the prediction task also influence the choice of an XAI method. Canha et al. [4] classify post-hoc XAI methods according to several criteria, including portability, scope, data type, and problem type. In their review, LIME and SHAP appear among the most widely discussed methods and are both described as model-agnostic and local methods applicable to tabular data, images, and text, as well as to classification and regression tasks [4]. Agarwal et al. [1] also identify LIME and SHAP as state-of-the-art feature attribution methods and include them in their benchmark of post-hoc explanation methods. This makes them particularly relevant for the present work, since the dataset is tabular, the prediction task is a regression task, and the objective is to compare explanations across several model architectures.

Finally, the form of the explanation must also be aligned with the target audience. Visual explanations, examples, simplified models, counterfactual explanations, or feature-relevance explanations do not answer the same interpretability need [2]. Barredo Arrieta et al. [2] distinguish several post-hoc explanation families, including visual explanations, local explanations, explanations by example, model simplification, and feature relevance methods. Feature relevance methods are particularly useful when the objective is to quantify the influence of each variable on the model output [2, 1]. In the context of this thesis, the objective is not to provide interactive explanations or counterfactual recommendations, but to understand which variables influence the prediction of the well-being score, both globally and locally. Other XAI methods also exist, such as permutation importance, Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) plots, counterfactual explanations, example-based explanations, and model simplification approaches. However, these methods were not retained in this thesis because the objective is to focus on feature-attribution explanations that can be applied consistently across the selected models and interpreted both at the global and local levels. For this reason, the following technical discussion focuses on LIME and SHAP, two widely used post-hoc methods that are suitable for comparing predictions across different machine learning models [1, 4].

1.5.2 Comparison and evaluation of explainability methods

The growing use of post-hoc explainability methods has raised an important methodological question: how can different explanations be compared and evaluated? This question is particularly relevant because XAI methods do not all produce the same type of output and do not necessarily answer the same interpretability need. Some methods provide feature attribution scores, others generate counterfactual examples, simplified rules, visual explanations, or representative examples [2, 4]. As a result, comparing XAI methods requires more than checking whether an explanation appears intuitive. It requires defining what property of the explanation is being assessed and in which context the explanation is used [7, 10, 1].

Doshi-Velez and Kim [7] propose a useful distinction between three levels of interpretability evaluation. Application-grounded evaluation relies on real users and real tasks, for example domain experts using explanations in an actual decision-making context. Human-grounded evaluation also involves human participants, but in simplified tasks designed to isolate specific aspects of explanation quality. Finally, functionally-grounded evaluation does not involve human subjects and instead relies on quantitative proxy metrics [7]. This last category is particularly relevant when human evaluation is difficult, costly, or

outside the scope of the study [4]. In this thesis, the comparison of XAI methods follows this functionally-grounded perspective, since the objective is to compare explanation patterns quantitatively across models and methods.

A first way to compare feature-attribution explanations is to examine whether they identify similar important variables. In this perspective, explanations can be transformed into ranked lists of features according to their absolute contribution. The overlap between the top-ranked features can then be measured. For instance, a top-k overlap indicates how many of the k most important features are shared by two explanations. This type of metric is close to the Feature Agreement metric described by Agarwal et al. [1], which measures the fraction of common features among the top-k features of two explanations. Such metrics are useful because they focus on the most influential variables, which are often the most relevant from an interpretability perspective [1, 4].

However, two explanations may select similar variables while ordering them differently. For this reason, rank-based metrics can be used to assess whether explanations not only identify the same features, but also assign them a similar relative importance. Agarwal et al. [1] discuss several metrics based on feature ranking, including Rank Agreement, Pairwise Rank Agreement, and Rank Correlation. In particular, Rank Correlation relies on Spearman's rank correlation coefficient to measure the agreement between two feature rankings [1]. This is relevant for the present thesis because Spearman rank correlation makes it possible to compare whether two models, or two XAI methods, organize feature importance in a similar way, even when the exact attribution values differ.

Beyond similarity between explanations, the literature also emphasizes the need to evaluate the reliability of explanation methods. One central criterion is faithfulness, sometimes also referred to as fidelity. Faithfulness assesses whether an explanation truly reflects the behavior of the predictive model [10, 1, 4]. In feature-attribution settings, this can be evaluated through perturbation-based metrics. The intuition is that if an explanation identifies a feature as important, then perturbing this feature should produce a larger change in the model output than perturbing features considered unimportant.

A common way to formalize this idea is the Prediction Gap on Important features (PGI), defined by Agarwal et al. [1] as follows:

$$PGI(x, f, e_x, k) = \mathbb{E}_{x' \sim \text{perturb}(x, e_x, \textit{top-}k)} \left[|\hat{y} - f(x')| \right], \quad \text{with } \hat{y} = f(x). \tag{1.6}$$

where x denotes the instance being explained, f the predictive model, e_x the explanation associated with x, and *top-k* the set of the k most important features according to the explanation. The perturbed instance x' is obtained by modifying the features identified as important while keeping the other features unchanged. A larger prediction gap suggests that the features highlighted by the explanation have a stronger influence on the model output, and may therefore indicate higher faithfulness. In the context of this thesis, this logic is adapted to a regression setting by measuring changes in the predicted well-being score rather than changes in classification probability.

Another important criterion is robustness, also referred to as stability. Robustness evaluates whether explanations remain stable when small perturbations are applied to the input data, especially when the model prediction remains close to the original prediction [10, 1]. A general formulation of this idea is the Relative Input Stability (RIS) metric:

$$RIS(x) = \frac{\|x\|_p}{\|e_x\|_p} \max_{x' \in N_x} \frac{\|e_x - e_{x'}\|_p}{\|x - x'\|_p}, \quad \text{with } \hat{y}_x \approx \hat{y}_{x'}. \tag{1.7}$$

where N_x denotes a local neighborhood around x, e_x is the explanation of the original instance, and $e_{x'}$ is the explanation obtained for the perturbed instance x'. This metric measures how much the explanation changes relative to the magnitude of the input perturbation. Lower values indicate more stable explanations. In this thesis, robustness is evaluated in a simplified local form by comparing the feature rankings obtained before and after perturbing the selected individual, using top-k overlap and Spearman rank correlation.

Therefore, the empirical analysis conducted in this thesis does not aim to provide an exhaustive benchmark of faithfulness and robustness metrics. Instead, it applies the underlying logic of these metrics to the available setting by testing whether an important feature perturbation affects the predicted score and

whether the resulting explanations remain stable after this perturbation.

Other evaluation criteria are also discussed in the XAI literature. Hedström et al. [10], for example, group XAI evaluation metrics into several categories, including faithfulness, robustness, localisation, complexity, randomisation, and axiomatic metrics. Agarwal et al. [1] also include fairness as an evaluation dimension, by examining whether explanation quality differs across majority and minority groups. Similarly, Canha et al. [4] emphasize the distinction between human-centered and functionally-grounded properties when evaluating XAI methods. These criteria show that XAI evaluation is multidimensional and that no single metric can fully determine whether an explanation is good [7, 10, 4]. The relevance of each criterion depends on the application, the available data, and the objective of the analysis [2, 4].

In the context of this thesis, the evaluation is therefore deliberately limited to criteria that are compatible with the available data and the methodological scope of the work. The comparison between models and explanation methods focuses first on the similarity of feature importance patterns, using top-k overlap and Spearman rank correlation. It then examines the reliability of explanations through faithfulness and robustness analyses. Fairness, human-grounded evaluation, and application-grounded evaluation remain important dimensions in the literature [7, 1, 4], but they are not treated as central empirical criteria here because they would require additional subgroup analyses, human participants, or domain-expert validation.

1.5.3 LIME (Local Interpretable Model-Agnostic Explanations)

Local Interpretable Model-Agnostic Explanations (LIME) [26] is a post-hoc explainability technique designed to interpret individual predictions produced by black-box models [9, 2, 14]. LIME belongs to the family of *local surrogate* methods [9, 32]: rather than approximating the global behaviour of a complex predictor, it fits a simple, human-interpretable model in the neighbourhood of a specific instance. The key assumption is that a model that is globally non-linear may still be *locally* approximated by an interpretable hypothesis class, typically a sparse linear model, thereby enabling instance-wise explanations [26, 32].

Problem formulation.

Let $f : \mathcal{X} \to R$ denote a trained black-box predictor (e.g., a neural network, random forest, or gradient-boosted ensemble) and let $x \in \mathcal{X}$ be the instance to be explained. LIME constructs an interpretable representation $z \in \mathcal{Z}$ of x (e.g., presence/absence of words for text, superpixel activations for images, or simplified binary/categorical encodings for tabular data) and seeks an explanation model g from an interpretable model class $\mathcal{G}$ (often sparse linear regressors). The LIME objective is commonly expressed as [26, 32]:

$$g^\star = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1.8}$$

where $\mathcal{L}(f, g, \pi_x)$ measures local fidelity (how well g approximates f in the vicinity of x), π_x is a proximity kernel that weights samples according to their distance to x, and $\Omega(g)$ is a complexity penalty that enforces interpretability (e.g., sparsity constraints) [26, 9].

Algorithmic workflow.

From a methodological perspective, LIME proceeds in the following steps [26, 32]:

1. **Perturbation sampling:** generate n synthetic observations by perturbing the interpretable representation z of the instance x (e.g., switching superpixels on/off for images, removing words for text, or sampling feature values for tabular data).
2. **Model querying:** obtain black-box outputs for each perturbed sample by querying f.
3. **Local weighting:** assign each perturbed sample a weight via a kernel function $\pi_x(\cdot)$ based on its distance to the original instance.
4. **Surrogate fitting:** fit an interpretable (often sparse) model g on the weighted dataset.
5. **Explanation extraction:** interpret g (e.g., feature coefficients or selected features) as the local explanation of the prediction $f(x)$.

Strengths.

LIME exhibits several properties that explain its wide adoption: (i) it is *model-agnostic* and can be applied to arbitrary predictors without requiring access to internal parameters; (ii) it provides *local*, instance-specific explanations that are useful for case-by-case decision support; and (iii) it is adaptable across data modalities (tabular, text, images) through suitable perturbation and interpretable representation choices [26].

Limitations: instability, locality sensitivity, and fidelity concerns.

Despite its success, multiple studies have highlighted methodological weaknesses of LIME, particularly regarding the reliability of the explanations under repeated runs and hyperparameter choices. The Bayesian extension proposed by Zhao et al.Zhao et al. [32] provides a detailed analysis and experimental evidence that these limitations can be severe in practice.

(1) Inconsistency in repeated explanations. A critical issue is that LIME can produce *inconsistent explanations* for the same instance when run multiple times. This behaviour primarily stems from randomness in the perturbation sampling process: when the number of perturbed samples n is limited (often necessary due to efficiency constraints), different random draws yield different surrogate fits and therefore different feature importance rankings. Zhao et al.Zhao et al. [32] demonstrate this phenomenon empirically and emphasise that the computational cost of LIME increases approximately linearly with n, making large n impractical for expensive models and strict latency requirements. They quantify this instability using rank-agreement metrics (e.g., Kendall's W) and show that agreement can be low for small sample sizes, precisely in settings where fast explanations are needed [32].

(2) Sensitivity to kernel width and the "locality definition" problem. LIME relies on a proximity kernel (commonly exponential) to define the local neighbourhood around x. A kernel width parameter controls how rapidly weights decay with distance; however, there is no generally accepted principled method to choose this parameter. As a result, explanations may change markedly as kernel settings vary. Zhao et al.Zhao et al. [32] formalise robustness to kernel width changes and propose an empirical robustness measure based on sampling pairs of kernel widths within a plausible interval and examining the induced variation in explanation vectors. Their results indicate that kernel sensitivity is not merely a theoretical concern but can materially alter the resulting explanations [32].

(3) Local surrogate fidelity and risk of misleading explanations. Because LIME fits an interpretable surrogate rather than directly exposing the causal mechanisms of f, high interpretability does not guarantee high *fidelity*. In other words, even if the surrogate is sparse and intuitive, it may not faithfully represent the black-box behaviour in the region of interest. Zhao et al.Zhao et al. [32] evaluate fidelity using causal-oriented metrics such as deletion and insertion tests and show that fidelity can be improved by integrating additional information beyond the perturbed-sample evidence typically used by standard LIME. This highlights that LIME explanations may be plausible yet incomplete or misleading when used to justify high-stakes decisions [32].

Implications and perspective.

Overall, LIME remains a foundational technique for local, model-agnostic interpretability and continues to serve as a key baseline in XAI research [26]. However, the evidence synthesised by Zhao et al. Zhao et al. [32] suggests that practitioners should exercise caution in sensitive deployments: explanation stability, robustness to locality definitions, and surrogate fidelity must be assessed (and, when possible, improved) before relying on LIME as a decision-support or auditing tool. This has motivated several extensions and refinements, including Bayesian formulations that aim to reduce sampling-induced variance and sensitivity to kernel settings by incorporating prior knowledge in a principled way [32].

1.5.4 SHAP (Shapley Additive Explanations)

SHAP is an explainability method grounded in cooperative game theory, and more specifically in Shapley values [19, 9, 2, 11]. In this framework, each feature is considered a player contributing to the final prediction, and its contribution is defined as its average marginal contribution across all possible coalitions of features [19, 2].

Formally, the Shapley value of a feature corresponds to the average difference in prediction observed when that feature is added to a subset of features, across all possible subsets. This formulation guarantees several fundamental theoretical properties: consistency, symmetry, additivity, and local accuracy [19, 9, 11]. These guarantees distinguish SHAP from many other explainability methods by providing a rigorous mathematical framework for contribution attribution [19, 2].

SHAP enables the production of local explanations similar to those of LIME, while also allowing contributions to be aggregated across observations to obtain global explanations of model behavior [19, 9, 11]. This dual capability is particularly useful for model auditing and understanding decision mechanisms. Moreover, optimized variants such as TreeSHAP make the computation of Shapley values tractable for certain families of models, particularly tree-based models such as XGBoost, which are widely used in mental health and well-being applications [19, 20, 3, 15].

Despite these advantages, SHAP also presents important limitations. Its computational cost can become prohibitive for complex or high-dimensional models when specific approximations are not available [19, 9, 11]. Additionally, the computation of Shapley values often relies on assumptions of feature independence, which are rarely satisfied in real-world clinical data and may bias the interpretation of contributions [19, 14, 22].

Another central issue concerns the nature of the explanations produced. Although mathematically well-defined, SHAP contributions remain correlational and should not be interpreted as causal relationships [14, 11]. Pendyala and Kim [22] emphasize that models with good predictive performance may still rely on irrelevant or clinically incoherent associations, despite seemingly robust SHAP explanations. This implies that SHAP should be used as a tool for critical analysis rather than as causal evidence of a model's reasoning [22, 31].

1.5.5 Model explainability and reliability

A key point raised by recent literature is that classical performance metrics are insufficient to assess the reliability of machine learning models. Pendyala and Kim [22] demonstrate that models with high accuracy may nevertheless rely on clinically irrelevant or incoherent variables, potentially leading to misleading predictions. XAI methods help reveal such problematic behaviors by shedding light on the internal mechanisms of models [22].

In a mental health context, Ntakolia et al. [20] illustrate how integrating XAI methods into an XGBoost-based pipeline enables the identification of key factors influencing emotional states while providing explanations that are actionable for experts. These works highlight that explainability is not only an interpretative tool but also a means of critical evaluation and model validation [20].

1.5.6 Current limitations and challenges of XAI methods

Despite their contributions, XAI methods present important limitations. The explanations produced are often local, approximate, and sensitive to methodological choices. Moreover, the lack of consensus on objective metrics to evaluate explanation quality remains a major challenge. Several authors emphasize that explainability should be considered a trade-off between model fidelity, human interpretability, and practical usefulness, rather than an absolute property [2, 31].

Chapter 2

Methodology

2.1 General context of the project

This work is part of the *Datagochi Santé* project, a collaborative research project carried out with several contributors, including researchers from UCLouvain and Université Laval.

The data used in the first study come from online questionnaires completed by a quota-based sample of Canadian adults recruited through the Léger Leo Web Panel. The initial questionnaire contained 280 items covering lifestyle behaviors, social engagement, sociodemographic characteristics, health-related information, and well-being. After excluding participants who failed attention checks, the final sample used for model development included 1,881 participants. The target variable was computed from participants' self-reported answers to the 14-item Mental Health Continuum-Short Form (MHC-SF), which produces a total well-being score. After filtering and feature selection, 20 questionnaire items were retained as predictors for the final model used to estimate well-being from lifestyle and contextual factors [30, 17].

2.2 Data and preprocessing

The present thesis uses the preprocessed dataset produced within the Datagotchi Santé project. As explained in the literature review, the original questionnaire was filtered and transformed into a reduced set of lifestyle and contextual predictors. The final modeling dataset is based on 20 selected questionnaire items, corresponding to 36 engineered features after encoding multicategory variables.

For the models tested in this thesis, preprocessing choices were included in the model selection process. Missing values could be handled either with a SimpleImputer or a KNNImputer, and feature scaling could be performed using either a StandardScaler or a MinMaxScaler. These preprocessing options were selected within the inner cross-validation loop, together with the model-specific hyperparameters. The selected preprocessing configuration for each fold is reported in the appendix.

2.3 Baseline models and initial state of the project

Before the integration of this work into the project, several regression models had already been evaluated using five-fold cross-validation. Performance was measured using the mean squared error (*Mean Squared Error*, MSE) and the mean absolute error (*Mean Absolute Error*, MAE).

The predicted target corresponds to the total well-being score, which ranges from 0 to 100. Therefore, the reported errors must be interpreted in relation to this scale. The MAE is expressed in the same unit as the target variable: for instance, an MAE of 14 means that, on average, the model's prediction differs from the true well-being score by approximately 14 points on a 100-point scale. The MSE, by contrast, is expressed in squared points and penalizes larger errors more strongly. For example, an MSE of approximately 321 corresponds to a root mean squared error of about 17.9 points, meaning that the typical prediction error is around 18 points when larger deviations are taken into account.

The models initially tested include:

- a *Random Regressor*, showing poor performance (MSE $\approx$ 1425, MAE $\approx$ 30.86),

- a *Mean Regressor*, serving as a simple baseline (MSE $\approx$ 474, MAE $\approx$ 17.83),
- a *Ridge Regressor*, achieving the best performance among the tested models (MSE $\approx$ 321, MAE $\approx$ 14.61),
- an *XGBoost Regressor*, with close but slightly lower performance (MSE $\approx$ 331, MAE $\approx$ 14.63).

The *Random Regressor* is a naive baseline model that produces predictions without learning a meaningful relationship between the input features and the target variable. For example, it can generate predictions randomly from the distribution of observed target values. The *Mean Regressor* is another simple baseline that always predicts the mean value of the target variable observed in the training set, independently of the input features.

The poor performance of the Random Regressor is therefore expected, because it is not designed to learn patterns from the data. Its role, together with the Mean Regressor, is to provide a reference point against which more advanced models can be compared. The fact that models such as the Ridge Regressor and XGBoost Regressor, both previously introduced and discussed in the literature review, obtain substantially lower errors indicates that they are able to capture meaningful relationships in the data and effectively learn from the selected features.

Figure 2.1 and Figure 2.2 summarize these average results over the cross-validation folds and motivate the use of the *Ridge Regressor* as the reference model for the remainder of the thesis.

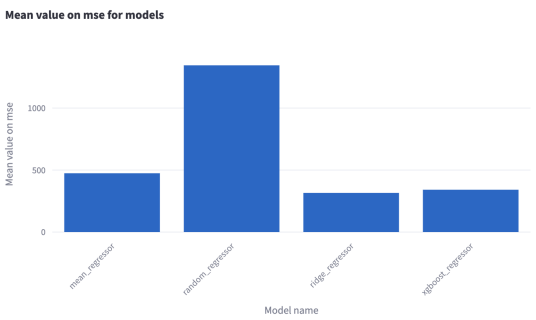

Figure 2.1: MSE

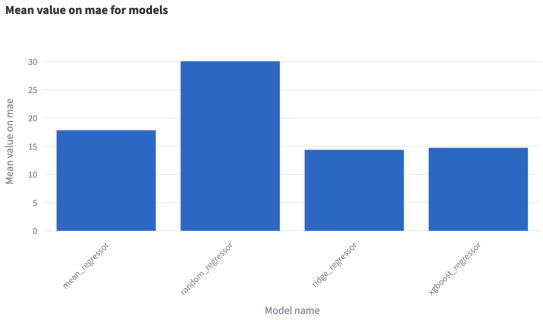

Figure 2.2: MAE

2.3.1 Ridge Regression as a Transparent Reference Model

The *Ridge Regressor* is retained as the reference model. As a regularized linear model, it offers intrinsic interpretability through the coefficients associated with the variables. The weights used in the project make it possible to quantify the directional association between each variable and the prediction of the well-being score.

For illustrative purposes, Table 2.1 presents the coefficients of the ridge regressor variables, which provides a first point of comparison for the XAI analyses conducted on the black-box models (the full questions text related to the variables are in the appendix).

Table 2.1: Ridge Regressor coefficients sorted in descending order.

Variable	Coef.	Variable	Coef.	Variable	Coef.	Variable	Coef.
sommeil_1	19.48	maladies_15	4.32	married_4.0	0.03	smoking	-3.32
autogestion_9	13.74	chronotype_3.0	3.11	travail_domaine_5	0.02	maladies_22	-3.41
act_friends	10.57	style_3.0	2.62	car_model_4.0	-0.10	consult_who_6	-3.69
quartier_domicile_3	10.37	maladies_20	2.21	nb_friends_dispo	-0.91	origines_ethniques_4.0	-4.54
act_volunteer	7.58	travail_domaine_1	1.61	chronotype_2.0	-1.02	travail_domaine_8	-4.74
act_nature_1	6.71	travail_domaine_6	0.72	maladies_16	-1.23	maladies_21	-5.68
issue_ai_data_3_4.0	6.21	maladies_18	0.70	consult_who_3	-1.87	consult_who_5	-5.78
style_2.0	5.80	Sex_2.0	0.47	travail_domaine_2	-2.42	maladies_19	-7.30
origines_ethniques_2.0	5.55	travail_domaine_9	0.19	married_5.0	-2.88	travail_domaine_3	-7.65
quartier_opportunite	5.28	partner_1.0	0.05	LatDec_3	-3.22	SoutSup_6	-10.92
style_8.0	4.35						

2.4 Search for high-performing black-box models

In the present thesis, additional models were tested to investigate whether more advanced machine learning approaches could improve predictive performance in the context of well-being prediction. The objective was to compare models of varying complexity with the baseline approaches previously introduced, including linear, non-linear, and ensemble-based methods.

The selection of models was guided by the reviewed literature, the characteristics of the dataset, and the need to ensure a representative comparison across different modeling paradigms. It should be noted, however, that it is not feasible to evaluate all existing models or all possible hyperparameter combinations. Therefore, the empirical analysis focuses on a subset of models considered most relevant to the research question.

All models were evaluated using the same five-fold cross-validation procedure. The target variable corresponds to the total well-being score derived from the Mental Health Continuum Short Form (MHC-SF). This score is bounded between 0 and 100, meaning that the prediction task consists of estimating a continuous well-being score on this scale. Predictive performance was primarily measured using the mean absolute error (Mean Absolute Error, MAE) and the mean squared error (Mean Squared Error, MSE). In addition, the coefficient of determination (R^2) was computed as a complementary metric, following the original Datagotchi Santé study, where MAE, MSE, and R^2 were used to evaluate model performance. Both metrics were computed for each fold, and their average across the five folds was used to compare the models.

For each outer fold, the model was trained on a training subset and evaluated on a held-out test subset. In order to avoid selecting hyperparameters directly on the test data, hyperparameter optimization was performed only within the training data. More specifically, for each outer fold, the corresponding training subset was further divided into three internal folds. Several candidate configurations were then tested, including different preprocessing choices and model-specific parameter values. For each candidate configuration, the model was trained and validated across the three internal folds, and the configuration giving the best average validation performance was retained. This selected configuration was then used to train the model on the full outer training subset and to evaluate it on the corresponding outer test fold.

This procedure corresponds to a nested cross-validation logic. The outer cross-validation loop provides an estimate of the model's generalization performance, while the inner cross-validation loop is used for model selection and hyperparameter tuning. As a result, the test fold remains independent from the hyperparameter selection process, which reduces the risk of overly optimistic performance estimates.

This also means that the hyperparameters were selected independently for each outer fold. Consequently, the same model can have different selected parameters from one fold to another, depending on the configuration that achieved the best internal cross-validation performance within the corresponding training subset. The final MSE and MAE reported for each outer fold therefore correspond to the performance obtained on the held-out test fold after this internal model selection procedure.

The same set of 20 selected questionnaire variables was used for all models. As explained previously, these 20 variables can generate a larger number of model features after encoding, particularly when categorical variables are transformed into numerical representations. The objective of using the same selected variables across all models is to ensure that the comparison focuses on model performance rather than on differences in input information.

Two preprocessing components were included in the model selection process: missing value imputation and feature scaling. Missing values could be handled either with a *SimpleImputer* or with a *KNNImputer*. The *SimpleImputer* replaces missing values with the mean of the corresponding feature computed on the training data. The *KNNImputer*, in contrast, estimates missing values using the values observed among the most similar individuals, based on a nearest-neighbor approach. Since no specific value of k was defined in the configuration, the default value of five neighbors was used.

Feature scaling could be performed using either a *StandardScaler* or a *MinMaxScaler*. The *StandardScaler* standardizes each feature by removing its mean and scaling it to unit variance, whereas the

MinMaxScaler rescales each feature to a fixed range, generally between 0 and 1.

2.5 Explainability analysis methodology

Following the predictive performance evaluation, an explainability analysis was conducted in order to examine how the selected models produced their predictions and to assess whether their explanations were consistent across models and methods. As discussed in the literature review, the objective of this thesis is not to evaluate all existing explainability methods exhaustively, but to focus on selected dimensions of explanation quality that are relevant to the present context and measurable with the available data and resources. More specifically, the analysis focuses on feature-attribution explanations and evaluates them through comparison, faithfulness, and robustness indicators.

Based on the literature reviewed in Section 1.5.1, LIME and SHAP were selected as the two post-hoc explainability methods used in this thesis. These methods were retained because they are widely used in the XAI literature, can be applied to tabular regression tasks, and provide feature-level explanations that are relatively intuitive to interpret. This choice is also consistent with the objective of this thesis, which is to compare the explainability of black-box models with that of a more interpretable linear baseline.

The explainability analysis was applied to a subset of four models. The Ridge Regressor was retained as an intrinsically interpretable baseline. The Stacking Regressor was retained because it obtained the best average predictive performance. CatBoost and LightGBM were also included because they obtained strong predictive results and represent tree-based black-box models. This selection makes it possible to compare explanations across models that differ in structure while remaining among the most relevant models from the predictive analysis.

2.5.1 Global explanation analysis

The first part of the explainability analysis focuses on global explanations. The objective is to examine whether the selected models rely on similar variables when predicting the well-being score at the dataset level. For this purpose, global SHAP explanations were computed for the four selected models.

For each model, the global importance of each feature was obtained by aggregating the absolute SHAP values across observations. The use of absolute values is important because the objective of this comparison is to measure the strength of each feature's contribution, regardless of whether the feature increases or decreases the predicted score. The resulting feature importances were then transformed into rankings, where the most influential feature received the highest rank.

Two complementary indicators were used to compare the global SHAP rankings between models. First, a top-k overlap was computed for the top 5 and top 10 most important features. This metric measures how many features are shared between the most influential variables identified by two models. Second, Spearman rank correlation was computed on the complete feature rankings. While the top-k overlap focuses only on the most important features, Spearman correlation evaluates whether the models rank all features in a similar order.

2.5.2 Local explanation analysis

The second part of the analysis focuses on local explanations. The selected individual was used as an illustrative local case rather than as a statistically representative observation of the full dataset. This individual was retained because the prediction was relatively low compared with the maximum possible well-being score and because the profile contained a clear and interpretable value for the feature `nb_friends_dispo`, equal to 0. This made it possible to analyze how the models explained a specific prediction and to later perform a simple and meaningful perturbation of this feature. Contrary to the global analysis, which examines the general behavior of the models across the dataset, the local analysis explains the prediction obtained for one selected individual. This makes it possible to compare how the models justify a specific predicted well-being score.

For this selected individual, local SHAP explanations were first computed for the four selected models. The SHAP contribution values indicate how each feature contributes to the individual prediction

compared with the model's reference prediction. Positive contributions increase the predicted well-being score, whereas negative contributions decrease it. However, for the purpose of comparing feature importance rankings across models, the SHAP values were ranked according to their absolute magnitude.

A similar local analysis was then conducted using LIME. Since LIME produces local feature contributions for a specific prediction, it was applied to the same selected individual and to the same four models. As with SHAP, the LIME contributions were ranked according to their absolute magnitude in order to compare the most influential features independently of the direction of their effect.

For both local SHAP and local LIME explanations, the same two comparison metrics were used. The top-k overlap was computed for the top 5 and top 10 local features, and Spearman rank correlation was computed on the complete feature rankings. This allowed the analysis to assess whether the models identified similar important features for the same individual and whether they ordered these features in a similar way.

2.5.3 Faithfulness analysis

After comparing the explanations across models, a faithfulness analysis was conducted. As explained in the literature review, faithfulness refers to the extent to which an explanation reflects the actual behavior of the predictive model. In this thesis, faithfulness was assessed through a local perturbation approach.

The analysis was conducted on the same selected individual as in the local SHAP and LIME analyses. For this selected individual, the feature `nb_friends_dispo` was modified from 0 to 1. This feature corresponds to the number of immediately available friends with whom the participant can talk frankly. The perturbation was therefore simple and interpretable: it represented a change from having no immediately available friend to having one immediately available friend. This feature was selected because it appeared among the most influential local features for several black-box models, making it relevant for a local faithfulness test. The four selected models were then applied again to the modified individual, and the new predicted well-being scores were compared with the original predictions.

The objective of this analysis was to observe whether modifying a feature identified as important by the explanation methods led to a change in the model prediction. Since the target variable is a well-being score, the faithfulness analysis was conducted by measuring the change in predicted score rather than a change in class probability. This analysis therefore provides a local indication of whether the explanations are coherent with the behavior of the models for the selected individual.

However, this analysis should be interpreted with caution. It does not constitute a complete evaluation of faithfulness across the entire dataset. It is limited to one selected observation and to the modification of one feature. The purpose is therefore to provide a focused local test rather than a general proof of faithfulness.

2.5.4 Robustness analysis

Finally, a robustness analysis was conducted in order to examine whether the local explanations remained stable after the same perturbation applied in the faithfulness analysis. More specifically, the value of `nb_friends_dispo` was changed from 0 to 1 for the selected individual, and SHAP and LIME explanations were recomputed for the four selected models. As discussed in the literature review, robustness refers to the stability of explanations when small changes are applied to the input data.

After modifying `nb_friends_dispo`, both SHAP and LIME explanations were recomputed for the four selected models. The feature rankings obtained after the modification were then compared with the original rankings. This comparison was performed separately for SHAP and LIME.

As in the previous comparison analyses, two indicators were used. First, the top-k overlap was computed for the top 5 and top 10 most influential features before and after the modification. Second, Spearman rank correlation was computed between the complete feature rankings before and after the perturbation. These indicators make it possible to assess whether the most important features remained similar and whether the overall ordering of feature importance was preserved.

This robustness analysis therefore evaluates whether the local explanations are stable for a slightly modified version of the same individual. As with the faithfulness analysis, the results must be interpreted as a local robustness test rather than as a general evaluation of explanation stability across the full dataset.

Chapter 3

Results

This chapter presents the empirical results obtained from the predictive and explainability analyses conducted in this thesis. First, the predictive performance of the different regression models is reported using MAE and MSE, based on the five outer cross-validation folds. The statistical comparison of these results is then presented in order to assess whether the observed differences between models are supported by statistical evidence. Finally, the chapter reports the explainability analyses conducted with SHAP and LIME, including global and local comparisons, as well as faithfulness and robustness analyses. The purpose of this chapter is therefore to present the observed results in a structured way, while their broader interpretation and implications are discussed in the following chapter.

3.1 Predictive performance results

The models retained for this empirical comparison were selected based on the literature reviewed earlier, the characteristics of the dataset, and the objective of comparing linear, non-linear, ensemble-based, and black-box approaches. For each model, the reported MAE and MSE correspond to the average performance obtained across the five outer cross-validation folds. The detailed fold-by-fold results, including the selected preprocessing steps and hyperparameters for each fold, are provided in the appendix 4.6.

Table 3.1 summarizes the average MAE and MSE obtained by each model. The models are ordered according to their average MAE, from the lowest to the highest value.

Table 3.1: Summary of average predictive performance across the five folds.

Model	**Average MAE**	**Average MSE**
Stacking Regressor	13.9615	300.0651
CatBoost Regressor	14.0363	301.2887
LightGBM Regressor	14.2625	313.2816
Ridge Regressor	14.3700	315.6651
Support Vector Regressor	14.4591	321.7933
Random Forest Regressor	14.7055	329.8884
XGBoost Regressor	14.7193	341.2881
ElasticNet Regressor	15.1224	345.3267
MLP Regressor	15.3425	366.0727
KNN Regressor	15.5407	370.9193
Mean Regressor	17.8316	474.2482
Random Regressor	30.8643	1387.9522

Overall, the lowest average error was obtained by the Stacking Regressor, with an average MAE of 13.9615 and an average MSE of 300.0651. The Stacking Regressor was defined with two base estimators: a CatBoost Regressor and a Ridge Regressor. The final estimator was selected within the inner cross-validation loop between a Ridge Regressor and a CatBoost Regressor. As a result, the selected final estimator could vary across outer folds. In most folds, Ridge was selected as the final estimator, while CatBoost was selected in one fold, as reported in the appendix 12. This configuration was selected because it achieved the best predictive performance among the different stacking configurations tested during the model selection process. The CatBoost Regressor obtained very close results, with an average

MAE of 14.0363 and an average MSE of 301.2887. These two models were followed by the LightGBM Regressor, the Ridge Regressor, and the Support Vector Regressor, all of which obtained average MAE values between approximately 14.3 and 14.6.

The MLP Regressor, ElasticNet Regressor, KNN Regressor, Random Forest Regressor, and XGBoost Regressor obtained intermediate results. Their average MAE values remained lower than those of the naive baselines, but higher than those of the best-performing models.

The Mean Regressor and the Random Regressor obtained substantially higher errors than the other models. The Mean Regressor reached an average MAE of 17.8316 and an average MSE of 474.2482, while the Random Regressor obtained the highest error values, with an average MAE of 30.8643 and an average MSE of 1387.9522. These results provide a reference point for assessing the extent to which the trained models reduce prediction error compared to naive approaches.

Although the Stacking Regressor achieved the best average performance, these descriptive results alone are not sufficient to conclude that it is statistically superior to the other models. The observed differences may partly result from variability across the cross-validation folds. For this reason, the next subsection presents the statistical comparison of predictive performance.

3.2 Statistical comparison of predictive performance

Although the average MAE and MSE values provide a first indication of the relative predictive performance of the models, they are not sufficient on their own to conclude that one model is statistically superior to another. In particular, the fact that the Stacking Regressor obtained the lowest average errors does not necessarily imply that it is significantly better than closely competing models such as CatBoost, LightGBM, or Ridge Regression. The observed ranking may partly result from variability across the cross-validation folds.

For this reason, statistical tests were applied to the fold-level predictive results in order to assess whether the observed differences between models were statistically significant. Since all models were evaluated on the same five outer cross-validation folds, the comparisons were performed on paired results rather than on independent observations. The objective of this subsection is therefore to determine whether the empirical ranking suggested by the average MAE and MSE values is statistically supported.

Following the statistical comparison procedures discussed in the literature, several non-parametric tests were considered. First, the Friedman test was applied to compare the models simultaneously across the five folds. This test provides an overall assessment of whether the models differ significantly in terms of predictive performance. When relevant, post-hoc pairwise comparisons were then used to further examine whether the Stacking Regressor was significantly better than the closest competing models.

It is important to note that these statistical comparisons are based on the five outer cross-validation folds. Therefore, the tests are not applied directly to the 1,881 individual observations in the dataset, but to five fold-level performance values for each model. As a result, the statistical analysis is based on a very small number of evaluation blocks. The results should therefore be interpreted with caution and considered as exploratory rather than as strong statistical evidence of superiority between models.

3.2.1 Friedman test

The Friedman test was first applied to compare the predictive performance of all models simultaneously. This non-parametric test is appropriate when several models are evaluated on the same cross-validation folds, because it takes into account the paired structure of the results. Instead of comparing only the average MAE values directly, the Friedman test ranks the models within each fold: the model with the lowest error receives the best rank, while the model with the highest error receives the worst rank. The test then evaluates whether the differences in average ranks across models are larger than would be expected by chance. The detailed fold-level ranks obtained by each model are reported in Appendix 13.

```
Friedman test
-------------
Statistic: 45.8705
p-value: 0.000003

Average ranks
-------------
Stacking Regressor          1.2
CatBoost Regressor          2.2
LightGBM Regressor          3.8
Ridge Regressor             5.0
Support Vector Regressor    5.5
XGBoost Regressor           6.0
ElasticNet Regressor        6.8
Random Forest Regressor     7.0
MLP Regressor               8.1
KNN Regressor               9.4
Mean Regressor             11.0
Random Regressor           12.0
dtype: float64
```

Figure 3.1: Friedman test results and average ranks obtained from the fold-level MAE values.

As shown in Figure 3.1, the Friedman test was applied to the fold-level MAE values of the twelve evaluated models. The null hypothesis states that all models have equivalent predictive performance, meaning that the observed differences in ranks are not statistically significant. The test returned a Friedman statistic of 45.8705 with a p-value of 0.000003. Since this p-value is lower than the conventional significance threshold of 0.05, the null hypothesis can be rejected. This indicates that there is a statistically significant difference in predictive performance between at least some of the evaluated models.

The average ranks provide additional information about the relative ordering of the models. Lower ranks correspond to better predictive performance, since they indicate lower MAE values across the folds. The Stacking Regressor obtained the best average rank (1.2), followed by the CatBoost Regressor (2.2), the LightGBM Regressor (3.8), and the Ridge Regressor (5.0). In contrast, the Mean Regressor and the Random Regressor obtained the worst average ranks, respectively 11.0 and 12.0. These results suggest that the Stacking Regressor was the most consistently well-ranked model across the five folds.

However, the Friedman test only indicates that at least one model differs significantly from the others. It does not identify which specific pairs of models are significantly different. Therefore, post-hoc pairwise comparisons are required to determine whether the Stacking Regressor is significantly better than the closest competing models, such as CatBoost, LightGBM, Ridge Regression, or the Support Vector Regressor.

3.2.2 Nemenyi post-hoc test

Since the Friedman test indicated a statistically significant difference between the models, a Nemenyi post-hoc test was applied to identify which pairwise differences were statistically significant. The Nemenyi test compares the average ranks of the models and determines whether the difference between two ranks is large enough to be considered statistically significant. As in the Friedman test, lower ranks correspond to better predictive performance, since they indicate lower MAE values across the folds.

In this analysis, the Nemenyi test was applied to the fold-level MAE values of the twelve evaluated models. Since the main objective was to assess whether the best-performing model was significantly better than the other models, Table 3.2 reports only the pairwise comparisons involving the Stacking Regressor, which obtained the best overall performance in terms of MAE. A complete pairwise comparison table, showing the significance of all two-by-two comparisons between models, is provided in the appendix in Table 14.

It should be noted that the Nemenyi post-hoc test was applied to only five evaluation blocks, corresponding to the five outer cross-validation folds. Therefore, the test was not based on the 1,881 individual observations of the dataset, but on five fold-level MAE values for each model. This is an important limitation, since the Nemenyi test is generally more appropriate when algorithms are compared across several independent datasets. In the present setting, the results should therefore be interpreted as exploratory and descriptive rather than as strong statistical evidence of superiority between models.

Table 3.2: Nemenyi post-hoc comparisons involving the Stacking Regressor.

Comparison	p**-value**	**Significant at 0.05**
Stacking vs Random Regressor	0.0001	Yes
Stacking vs Mean Regressor	0.0010	Yes
Stacking vs KNN Regressor	0.0169	Yes
Stacking vs MLP Regressor	0.1011	No
Stacking vs Random Forest Regressor	0.3127	No
Stacking vs ElasticNet Regressor	0.3683	No
Stacking vs XGBoost Regressor	0.6192	No
Stacking vs Support Vector Regressor	0.7685	No
Stacking vs Ridge Regressor	0.8835	No
Stacking vs LightGBM Regressor	0.9929	No
Stacking vs CatBoost Regressor	1.0000	No

The results show that the Stacking Regressor was significantly better than the weakest models in the comparison. More specifically, the difference was statistically significant between the Stacking Regressor and the Random Regressor ($p = 0.0001$), the Mean Regressor ($p = 0.0010$), and the KNN Regressor ($p = 0.0169$). These results indicate that the best-ranked model clearly outperformed the naive baselines and the KNN Regressor.

However, the Nemenyi test did not show a statistically significant difference between the Stacking Regressor and the closest competing models. The comparisons with CatBoost ($p = 1.0000$), LightGBM ($p = 0.9929$), Ridge Regression ($p = 0.8835$), Support Vector Regression ($p = 0.7685$), XGBoost ($p = 0.6192$), ElasticNet ($p = 0.3683$), Random Forest ($p = 0.3127$), and MLP ($p = 0.1011$) were not statistically significant at the 0.05 threshold.

This result is also consistent with the critical difference analysis. With twelve models and five folds, the critical difference was estimated at 7.4522 (see 1.3.2). The rank difference between the Stacking Regressor and CatBoost was only 1.0, and the rank differences with LightGBM, Ridge Regression, Support Vector Regression, and XGBoost were also below the critical difference threshold. In contrast, the rank differences between the Stacking Regressor and KNN, Mean Regressor, and Random Regressor exceeded the critical difference threshold.

Overall, the Nemenyi post-hoc test indicates that the Stacking Regressor achieved the best average rank. However, given that the comparison is based on only five folds from a single dataset, this result should not be interpreted as strong statistical evidence. The test mainly suggests that the Stacking Regressor performs better than the weakest models, while its advantage over the strongest competing models remains uncertain.

3.2.3 Pairwise Wilcoxon signed-rank tests with Holm correction

The Nemenyi post-hoc test provides a global pairwise comparison based on average ranks. However, because the main objective of this analysis is to assess whether the Stacking Regressor is significantly better than the closest competing models, additional pairwise Wilcoxon signed-rank tests were conducted. This test is appropriate because the models were evaluated on the same five outer cross-validation folds, meaning that the fold-level MAE values are paired. However, the very small number of paired observations must be emphasized. Since only five fold-level MAE values are available for each model, the Wilcoxon signed-rank tests have limited statistical power. Consequently, non-significant results should not be interpreted as evidence that the models are equivalent.

The Wilcoxon signed-rank tests were applied to compare the Stacking Regressor with the four closest competing models according to the average MAE and average rank results: CatBoost, LightGBM, Ridge Regression, and Support Vector Regression. A one-sided alternative hypothesis was used, testing whether the Stacking Regressor obtained lower MAE values than each compared model.

Since several pairwise comparisons were performed, Holm correction was applied to the resulting p-values in order to control the risk of false positive conclusions due to multiple testing. This correction is

necessary because performing several statistical tests increases the probability of obtaining a significant result by chance. The Holm procedure works by ordering the raw p-values from the smallest to the largest and then comparing them to progressively less restrictive significance thresholds. In practice, the smallest p-value is evaluated most strictly because it has the highest risk of being a false positive among the multiple comparisons. If a comparison remains significant after this adjustment, it can be considered more robust than if it were assessed using the raw p-value alone. In the present setting, however, the Holm correction should also be interpreted with caution. Since the number of observations is already very small, applying a multiple-comparison correction further reduces the ability of the tests to detect differences between models. Therefore, the corrected results are reported mainly for transparency and should not be overinterpreted.

In addition to the targeted comparisons presented in this section, a complete pairwise Wilcoxon signed-rank comparison matrix was also produced and is reported in Appendix 15. This appendix table compares each model with every other model two by two and indicates whether the corresponding difference remains significant after Holm correction.

Table 3.3: Pairwise Wilcoxon signed-rank tests comparing the Stacking Regressor with the closest competing models.

Comparison	**Raw p-value**	**Holm-corrected p-value**	**Folds won**	**Significant**
Stacking vs CatBoost	0.0625	0.1250	4/5	No
Stacking vs LightGBM	0.0313	0.1250	5/5	No
Stacking vs Ridge	0.0313	0.1250	5/5	No
Stacking vs Support Vector Regressor	0.0313	0.1250	5/5	No

As shown in Table 3.3, the Stacking Regressor obtained lower MAE values than CatBoost in four out of five folds, and lower MAE values than LightGBM, Ridge Regression, and Support Vector Regression in all five folds. Before correction, the comparisons with LightGBM, Ridge Regression, and Support Vector Regression produced raw p-values of 0.0313, which are below the conventional significance threshold of 0.05. The comparison with CatBoost produced a raw p-value of 0.0625, which is slightly above this threshold.

However, after applying Holm correction, all corrected p-values were equal to 0.1250. Consequently, none of the pairwise differences remained statistically significant at the 0.05 level. These results indicate that, although the Stacking Regressor achieved lower MAE values than the closest competing models in most or all folds, the statistical evidence is not sufficient to conclude that it is significantly better after correction for multiple comparisons.

These results should therefore be interpreted descriptively. The Stacking Regressor obtained lower MAE values than the closest competing models in most or all folds, which suggests a promising predictive pattern. However, because the comparison is based on only five fold-level values, the statistical evidence remains too limited to conclude that the Stacking Regressor is significantly superior to CatBoost, LightGBM, Ridge Regression, or Support Vector Regression.

3.3 XAI

Although the Stacking Regressor obtained the lowest mean MAE, the statistical tests showed that its performance was not significantly different from several other models after Holm correction. Therefore, the explainability analysis does not assume that the Stacking Regressor is statistically superior to all competing models. Instead, the analysis focuses on a representative subset of models.

The Ridge Regressor is retained as an intrinsically interpretable baseline. The Stacking Regressor is retained because it achieved the best mean predictive performance. In addition, Light GBM and CatBoost regressor are included in order to examine whether similar predictive performance leads to similar explanations across model architectures.

3.3.1 Global comparison of SHAP explanations

In order to compare the global explanations produced by the selected models, two simple indicators were computed from the SHAP feature rankings. The objective of this comparison is to assess whether the models rely on similar variables when predicting the well-being score. This analysis is particularly useful because models with similar predictive performance may still use the input variables differently.

First, a top-k overlap was computed. This indicator measures how many features are shared between the most important features identified by two models. In this thesis, the overlap was computed for the top 5 and top 10 SHAP features. A high overlap indicates that two models identify similar predictors as important, whereas a lower overlap suggests that they rely on different patterns of information.

Second, a Spearman rank correlation was computed between the SHAP feature rankings of each pair of models. This metric evaluates whether two models rank the features in a similar order. Unlike the top-k overlap, which focuses only on the most important variables, the Spearman correlation takes the complete feature ranking into account.

Table 3.4: Top-k overlap between SHAP feature rankings of selected models.

Model comparison	**Top-5 overlap**	**Top-10 overlap**
Stacking – CatBoost	4/5	9/10
Stacking – LightGBM	4/5	9/10
Stacking – Ridge	3/5	8/10
CatBoost – LightGBM	4/5	10/10
CatBoost – Ridge	4/5	8/10
LightGBM – Ridge	3/5	8/10

Table 3.4 shows that the selected models rely on highly similar variables at the global level. The strongest similarity is observed between CatBoost and LightGBM, which share all of their top 10 SHAP features. The Stacking Regressor also shows a strong overlap with both CatBoost and LightGBM, sharing 9 out of 10 top features with each of them. This suggests that the Stacking Regressor relies on a set of important variables that is very close to those identified by the two gradient boosting models.

The comparison with the Ridge Regressor shows slightly lower, but still substantial, overlap. Ridge shares 8 out of its top 10 features with Stacking, CatBoost, and LightGBM. This indicates that the linear reference model identifies several of the same important predictors as the black-box models, even though the exact ranking of these predictors may differ.

Table 3.5: Spearman rank correlation between complete SHAP feature rankings.

Model comparison	**Spearman ρ**
Stacking – CatBoost	0.956
Stacking – LightGBM	0.963
Stacking – Ridge	0.807
CatBoost – LightGBM	0.969
CatBoost – Ridge	0.756
LightGBM – Ridge	0.801

The Spearman correlations confirm the patterns observed in the top-k overlap analysis. The strongest agreement is found between CatBoost and LightGBM, with a Spearman correlation of 0.969. The Stacking Regressor also presents very high correlations with both LightGBM and CatBoost, with correlations of 0.963 and 0.956 respectively. These results suggest that the black-box models rank the features in a very similar way.

The correlations involving the Ridge Regressor are lower, although they remain positive and relatively high. The correlation between Ridge and Stacking is 0.807, while the correlations with CatBoost and LightGBM are 0.756 and 0.801 respectively. This indicates that the Ridge Regressor identifies several similar important variables, but that its complete feature ranking differs more from the rankings produced by the black-box models.

Overall, these results suggest that the selected models rely on a common group of important predictors, including variables related to sleep, self-management, social support, volunteering, social activities, and neighborhood context. However, the slightly lower agreement between the Ridge Regressor and the black-box models suggests that non-linear models may capture different patterns of importance for some variables. For example, *nb_friends_dispo* is highly ranked by the black-box models, whereas it does not appear among the top 10 SHAP features of the Ridge Regressor.

It should be noted that the top-k overlap is calculated only on the variables present in the top 5 or top 10 SHAP features of each model. In contrast, the Spearman rank correlation is calculated on the complete SHAP feature ranking. These two indicators are therefore complementary: the top-k overlap focuses on the most influential variables, while the Spearman correlation evaluates the similarity of the overall ranking. The complete feature rankings used for this comparison are provided in Appendix 16.

3.3.2 Local comparison of SHAP explanations

After comparing the models at the global level, a local SHAP analysis was conducted on one selected test observation. The objective of this analysis is to assess whether the models identify similar influential features when explaining the prediction of the same individual. Unlike the global SHAP analysis, which summarizes feature importance across the whole dataset, the local analysis focuses on one specific prediction.

For this comparison, the SHAP values were ranked according to their absolute magnitude. Therefore, the ranking reflects the strength of each feature contribution to the prediction, regardless of whether the contribution is positive or negative. The selected individual is described in Appendix Table 17, while the complete local SHAP rankings are reported in Appendix Table 18.

As in the global comparison, two indicators were used: the top-k overlap and the Spearman rank correlation. The top-k overlap measures how many features are shared between the most influential features identified by two models. In this thesis, the overlap was computed for the top 5 and top 10 local SHAP features. The Spearman correlation evaluates whether two models rank the complete set of features in a similar order.

Table 3.6: Top-k overlap between local SHAP feature rankings of selected models.

Model comparison	**Top-5 overlap**	**Top-10 overlap**
Stacking – CatBoost	4/5	8/10
Stacking – LightGBM	3/5	7/10
Stacking – Ridge	4/5	7/10
CatBoost – LightGBM	4/5	7/10
CatBoost – Ridge	3/5	7/10
LightGBM – Ridge	2/5	7/10

Table 3.6 shows that the models identify several common influential features for the selected individual, although the agreement is weaker than in the global comparison. The strongest local overlap is observed between the Stacking Regressor and CatBoost, which share four of their top 5 features and eight of their top 10 features. This suggests that the local explanation produced by the Stacking Regressor remains close to the explanation produced by CatBoost.

CatBoost and LightGBM also show a strong top-5 overlap, with four shared features among their five most influential local predictors. This indicates that the two gradient boosting models identify a similar group of dominant features for this individual. However, their top-10 overlap is lower than in the global analysis, which suggests that local explanations may be more sensitive to model-specific behavior.

The comparison with the Ridge Regressor shows a more mixed pattern. Ridge shares four of its top 5 features with Stacking, but only two of its top 5 features with LightGBM. This indicates that the linear reference model partially agrees with the black-box models on the most influential local features, but

does not rank all feature contributions in the same way. This difference is also visible in the complete local rankings reported in Appendix Table 18.

Table 3.7: Spearman rank correlation between complete local SHAP feature rankings.

Model comparison	**Spearman ρ**
Stacking – CatBoost	0.713
Stacking – LightGBM	0.720
Stacking – Ridge	0.540
CatBoost – LightGBM	0.844
CatBoost – Ridge	0.697
LightGBM – Ridge	0.742

The Spearman correlations reported in Table 3.7 confirm the patterns observed in the top-k overlap analysis. The highest correlation is observed between CatBoost and LightGBM, with a Spearman correlation of 0.844. This means that the two gradient boosting models produce the most similar complete local feature ranking for the selected individual.

The Stacking Regressor also shows positive correlations with CatBoost and LightGBM, with correlations of 0.713 and 0.720 respectively. These values indicate a moderate to strong similarity in the ordering of local feature importance. However, these correlations are lower than those observed in the global SHAP comparison, suggesting that model explanations are more aligned at the dataset level than for a single individual prediction.

The lowest correlation is observed between Stacking and Ridge, with a Spearman correlation of 0.540. This indicates that the linear model differs more substantially from the ensemble model in the way it ranks local feature importance. Although Ridge shares some important features with the other models, its complete local ranking is less similar to the rankings produced by the black-box models.

Overall, the local SHAP comparison shows that the selected models identify a partially common set of influential features for the same individual. The strongest consistency is observed between the black-box models, especially CatBoost and LightGBM, while the Ridge Regressor appears more distinct in its complete ranking. Compared with the global SHAP analysis, the local results show lower agreement between models. This suggests that the models rely on similar general predictors at the global level, but that their explanations may diverge more when explaining a specific individual prediction.

Although the present analysis focuses mainly on feature rankings rather than on the detailed contribution values, the local SHAP values are reported in Appendix 19 for completeness. These results should be interpreted with caution. The local analysis concerns only one selected observation and therefore cannot be generalized to the whole dataset.

3.3.3 Local comparison of LIME explanations

After the local SHAP comparison, a similar analysis was conducted using LIME. Since LIME is a local explanation method, the comparison focuses on the same selected individual as in the previous subsection. The objective is to assess whether the four models identify similar influential features when explanations are produced with another XAI method.

For this comparison, the LIME contributions were ranked according to their absolute magnitude. A lower rank therefore indicates a stronger local contribution to the prediction, regardless of whether the feature increases or decreases the predicted well-being score. The complete LIME rankings are reported in Appendix Table 20, while the detailed LIME contribution values and predicted scores are reported in Appendix Table 21.

As for the SHAP analyses, two indicators were computed: the top-k overlap and the Spearman rank correlation. The top-k overlap compares the features appearing among the most important predictors, while the Spearman correlation evaluates the similarity of the complete feature rankings.

Table 3.8: Top-k overlap between local LIME feature rankings of selected models.

Model comparison	**Top-5 overlap**	**Top-10 overlap**
Stacking – CatBoost	4/5	8/10
Stacking – LightGBM	4/5	8/10
Stacking – Ridge	2/5	8/10
CatBoost – LightGBM	4/5	9/10
CatBoost – Ridge	1/5	7/10
LightGBM – Ridge	1/5	7/10

Table 3.8 shows that the local LIME explanations are relatively consistent between the black-box models. The strongest agreement is observed between CatBoost and LightGBM, which share four of their top 5 features and nine of their top 10 features. The Stacking Regressor also shows a strong overlap with CatBoost and LightGBM, sharing eight of the top 10 features with both models.

The comparison with Ridge is more mixed. Ridge shares only two of its top 5 features with Stacking, and only one with CatBoost and LightGBM. However, the top-10 overlap remains relatively high, with seven or eight shared features depending on the comparison. This suggests that Ridge identifies several similar important features, but gives them a different priority in the local ranking.

Table 3.9: Spearman rank correlation between complete local LIME feature rankings.

Model comparison	**Spearman ρ**
Stacking – CatBoost	0.907
Stacking – LightGBM	0.871
Stacking – Ridge	0.745
CatBoost – LightGBM	0.887
CatBoost – Ridge	0.553
LightGBM – Ridge	0.510

The Spearman correlations reported in Table 3.9 confirm this pattern. The Stacking Regressor is strongly correlated with CatBoost and LightGBM, with correlations of 0.907 and 0.871 respectively. CatBoost and LightGBM also present a high correlation of 0.887, indicating that the two gradient boosting models produce similar complete LIME rankings for the selected individual.

The correlations involving Ridge are lower, especially with CatBoost and LightGBM. This indicates that the linear model orders the local LIME feature contributions differently from the black-box models. However, the positive correlations show that the rankings are not completely divergent.

Overall, the local LIME comparison leads to a conclusion similar to the local SHAP analysis. The black-box models, especially CatBoost, LightGBM, and Stacking, tend to identify a common set of influential features for the selected individual. Ridge remains more distinct in its ranking, particularly among the most influential features. These results suggest that local explanations are partly consistent across models, but that the exact ranking of feature importance depends on both the predictive model and the explanation method used.

As with the previous local analysis, these results should be interpreted with caution. The analysis concerns only one selected observation and therefore cannot be generalized to the whole dataset. All the LIME value for the observation are available in the appendix Table 19 if needed.

3.3.4 Faithfulness of local XAI models

The faithfulness analysis was conducted on the same selected individual as in the previous local SHAP and LIME analyses. For this individual, the value of `nb_friends_dispo` was modified from 0 to 1. This corresponds to a change from reporting no immediately available friend to reporting one immediately available friend. This perturbation was chosen because `nb_friends_dispo` appeared as an important local feature for several black-box models.

After modifying *nb_friends_dispo*, the four selected models were run again on the modified individual. The objective of this subsection is to report the observed change in the predicted well-being score after the modification of this feature. The interpretation of these results in relation to the faithfulness of the explanation methods is discussed in the discussion section.

Table 3.10: Effect of modifying *nb_friends_dispo* on the predicted well-being score.

Model	**Original prediction**	**Modified prediction**	**Prediction change**
CatBoost	42.143	48.612	+6.469
LightGBM	42.046	49.266	+7.220
Ridge	50.051	49.206	-0.845
Stacking	42.866	46.919	+4.053

Table 3.10 reports the predicted scores obtained before and after modifying *nb_friends_dispo*. For CatBoost, the predicted score increased from 42.143 to 48.612, corresponding to a change of +6.469 points. For LightGBM, the predicted score increased from 42.046 to 49.266, corresponding to a change of +7.220 points. For the Stacking Regressor, the predicted score increased from 42.866 to 46.919, corresponding to a change of +4.053 points. In contrast, the Ridge Regressor showed a smaller change, with the predicted score decreasing from 50.051 to 49.206, corresponding to a change of -0.845 points.

These results show that the modification of *nb_friends_dispo* led to different changes in predicted scores depending on the model. The largest changes were observed for LightGBM and CatBoost, followed by the Stacking Regressor. The smallest change was observed for the Ridge Regressor.

This analysis is limited to one selected individual and to the modification of one feature only. Therefore, the results reported in this subsection should be interpreted as a local perturbation analysis rather than as a general evaluation of faithfulness across the entire dataset.

3.3.5 Robustness of local XAI models

The robustness analysis was conducted using the same modified individual as in the previous faithfulness analysis. After modifying *nb_friends_dispo*, SHAP and LIME explanations were recomputed for the four selected models. The objective of this subsection is to report how similar the local explanations remained after the modification of this feature.

The complete local contribution values obtained after modifying *nb_friends_dispo* are reported in the appendix. The detailed LIME values are provided in Table 23, while the detailed local SHAP values are provided in Table 24.

For each model and each explanation method, the feature rankings obtained before and after the modification were compared. As in the previous XAI comparisons, two indicators were used. First, the top-k overlap was computed for the top 5 and top 10 most influential features. Second, the Spearman rank correlation was computed between the complete feature rankings before and after modification.

Table 3.11: Robustness of local SHAP explanations after modifying *nb_friends_dispo*.

Model	**Top-5 overlap**	**Top-10 overlap**	**Spearman ρ**
CatBoost	5/5	10/10	0.965
LightGBM	5/5	8/10	0.915
Ridge	5/5	9/10	0.939
Stacking	5/5	9/10	0.938

Table 3.11 reports the comparison between the original local SHAP rankings and the local SHAP rankings obtained after modifying *nb_friends_dispo*. For all four models, the top-5 overlap was equal to 5/5, meaning that the five most influential SHAP features remained the same after the modification. The top-10 overlap was also high. CatBoost obtained a top-10 overlap of 10/10, while Ridge and Stacking both obtained an overlap of 9/10. LightGBM showed a slightly lower top-10 overlap, with 8 out of 10

features remaining common.

The Spearman rank correlations were also high for the four models. CatBoost obtained the highest correlation, with a value of 0.965. Ridge and Stacking obtained similar values, respectively 0.939 and 0.938. LightGBM obtained a Spearman correlation of 0.915. These values indicate that the complete SHAP feature rankings before and after modification remained highly similar.

Table 3.12: Robustness of local LIME explanations after modifying *nb_friends_dispo*.

Model	**Top-5 overlap**	**Top-10 overlap**	**Spearman ρ**
CatBoost	5/5	10/10	1.000
LightGBM	5/5	10/10	0.999
Ridge	5/5	10/10	1.000
Stacking	5/5	10/10	0.999

Table 3.12 reports the same comparison for the local LIME explanations. For all four models, the top-5 overlap was equal to 5/5 and the top-10 overlap was equal to 10/10. This means that the same features appeared among the five and ten most influential LIME features before and after the modification.

The Spearman rank correlations were also very close to 1. CatBoost and Ridge obtained correlations of 1.000, while LightGBM and Stacking obtained correlations of 0.999. These results show that the complete LIME rankings before and after modification were almost identical for the selected individual.

Overall, the robustness results show that the local explanations remained highly similar after modifying *nb_friends_dispo*. This pattern is observed for both SHAP and LIME. The LIME rankings changed very little, with complete top-10 overlap and Spearman correlations close to 1 for all models. The SHAP rankings also remained highly similar, although small differences appeared in the top-10 rankings of LightGBM, Ridge, and Stacking.

As in the previous subsection, this analysis is limited to one selected individual and to the modification of one feature only. Moreover, the perturbation affected the predicted score differently depending on the model, as reported in the faithfulness analysis. Therefore, the results presented here should be interpreted as a local robustness analysis of the explanations for this specific case, rather than as a general evaluation of explanation robustness across the full dataset.

Chapter 4

Discussion

4.1 Interpretation of predictive performance

The first objective of this thesis was to examine whether more complex machine learning models could improve the prediction of well-being compared to the linear reference model. The results show that the Stacking Regressor achieved the lowest average prediction error, with an MAE of 13.9615 and an MSE of 300.0651. It was followed very closely by the CatBoost Regressor, with an MAE of 14.0363 and an MSE of 301.2887, and by the LightGBM Regressor, with an MAE of 14.2625 and an MSE of 313.2816. The Ridge Regressor, used as the more transparent reference model, also remained highly competitive, with an MAE of 14.3700 and an MSE of 315.6651. These results indicate that the best-performing models were all able to reduce prediction error compared to the naive baselines, but that the differences between the strongest models were relatively small.

This result suggests that non-linear and ensemble-based models can bring a predictive advantage in the context of well-being prediction, but that this advantage remains limited. The Stacking Regressor obtained the best average performance, which is consistent with the idea that combining complementary models can help capture complex relationships in the data. In this case, the stacking approach may have benefited from the combination of a tree-based model, able to capture non-linear effects, and a linear final estimator, able to combine the information in a more regularized way. However, the improvement over CatBoost, LightGBM, and Ridge Regression was not large enough to clearly separate the Stacking Regressor from these models on predictive performance alone.

The statistical tests reinforce this cautious interpretation, but their own limitations must also be emphasized. The Friedman, Nemenyi, and Wilcoxon analyses were applied to the five outer cross-validation folds. Therefore, the statistical comparisons were based on only five fold-level MAE values for each model, rather than on several independent datasets. This strongly limits the statistical power of the tests and makes significant results difficult to obtain, especially between models with close average performance. Consequently, the absence of significant differences between the Stacking Regressor and the strongest competing models should not be interpreted as evidence that these models are equivalent. Rather, it indicates that the available fold-level evidence is insufficient to support a strong statistical superiority claim.

For this reason, the predictive results should mainly be interpreted descriptively. The Stacking Regressor obtained the lowest average MAE and MSE, followed very closely by CatBoost and LightGBM. This is an interesting result because it suggests that ensemble and boosting approaches may be promising for well-being prediction. However, the differences between the strongest models remain small, and the Ridge Regressor remains highly competitive despite its simpler and more interpretable structure.

These findings are important because they nuance the interpretation of the predictive results. From a purely descriptive point of view, the Stacking Regressor appears to be the best model. However, from a statistical point of view, the evidence is not sufficient to conclude that it is clearly superior to the other strongest models. Therefore, the results should not be interpreted as showing a strong domination of black-box or ensemble models over the linear baseline. Rather, they suggest that several models reach a similar level of predictive performance on this dataset.

This point is particularly relevant in the context of well-being prediction. Since the target variable is a well-being score on a 0–100 scale, an MAE around 14 means that the model prediction differs from the observed score by approximately 14 points on average. Although this level of error is clearly better than the Mean Regressor and the Random Regressor, it also shows that the prediction of well-being remains a difficult task. This is not surprising, given that well-being is a multidimensional and subjective construct influenced by psychological, social, lifestyle, and contextual factors. Even with selected predictors and advanced models, a substantial part of the individual variability in well-being remains difficult to predict.

The comparison with the Ridge Regressor is especially informative. Ridge Regression is much simpler and more transparent than the black-box models, yet its performance remains close to that of the best-performing models. This suggests that the relationships captured by the selected features may be partly approximated by a linear model, or at least that the additional complexity of non-linear models only provides a modest improvement in this setting. Consequently, the choice of a black-box model cannot be justified solely by predictive performance. In a sensitive domain such as well-being, where model outputs may influence interpretation or decision-making, a small gain in accuracy must be weighed against the loss of intrinsic interpretability.

Overall, the predictive performance results indicate that black-box models, and especially the Stacking Regressor, can slightly improve well-being score prediction. However, this improvement is not statistically strong enough to establish a clear superiority over the best competing models. This finding motivates the explainability analysis conducted in the following sections: if black-box models are to be considered relevant in this context, their use must be supported not only by predictive performance, but also by the ability to provide explanations that are coherent, stable, and interpretable.

4.2 Explainability of black-box models compared to the linear baseline

The second objective of this thesis was to examine whether black-box models can be made explainable in the context of well-being prediction, and how their explanations compare with those of a more transparent linear model. Since the predictive performance results did not show a clear statistical superiority of the best black-box models over the Ridge Regressor, the explainability analysis plays a central role in determining whether the additional model complexity remains justified.

4.2.1 Global explanations

At the global level, the SHAP analysis shows a strong agreement between the selected black-box models. CatBoost and LightGBM share all their top 10 most important features, while the Stacking Regressor shares 9 out of its top 10 features with both CatBoost and LightGBM. This suggests that these models rely on a very similar group of predictors when estimating the well-being score. The Spearman rank correlations confirm this result, with very high correlations between CatBoost and LightGBM, as well as between the Stacking Regressor and the two gradient boosting models.

This global consistency is important because it suggests that the black-box models do not base their predictions on completely different or unstable patterns. On the contrary, they tend to identify a common set of relevant variables, including sleep quality, available friends, self-management, neighborhood perception, volunteering, social activities, and social support. These variables are also coherent with the multidimensional nature of well-being discussed in the literature review, since they reflect psychological, social, lifestyle, and contextual dimensions.

The comparison with the Ridge Regressor is more nuanced. Ridge shares several important variables with the black-box models, especially in the top 10 features, which suggests that the linear baseline and the black-box models partly rely on similar information. However, the complete rankings are less similar for Ridge than for the black-box models. This indicates that the linear model does not prioritize all variables in the same way. For example, *nb friends dispo* is highly ranked by the black-box models, whereas it is much less important in the Ridge ranking. This may suggest that the black-box models capture non-linear patterns or interactions that are not represented in the same way by the linear model.

4.2.2 Local explanations

The local SHAP and LIME analyses provide a more detailed view of how the models explain the prediction of one selected individual. Compared with the global analysis, the agreement between models is lower at the local level. This result is expected, because local explanations focus on a single prediction and are therefore more sensitive to the specific characteristics of the selected individual and to the internal behavior of each model.

For SHAP, the strongest local agreement is observed between CatBoost and LightGBM, while the Stacking Regressor also remains relatively close to these two models. Ridge, however, differs more clearly from the black-box models in its complete local ranking. This shows that, even when models reach similar predictive performance, they may justify an individual prediction through different feature importance patterns.

The LIME results follow a similar logic. The black-box models, especially CatBoost, LightGBM, and the Stacking Regressor, show strong agreement in their local feature rankings. Ridge again appears more distinct, particularly among the most influential features. This suggests that the difference between the linear model and the black-box models is not only visible in predictive structure, but also in the way individual predictions are explained.

4.2.3 Comparison between global and local explainability

Overall, the results show that black-box models are more consistent at the global level than at the local level. Globally, the models identify highly similar important predictors, which supports the idea that their general behavior can be interpreted in a coherent way. Locally, however, the explanations become more model-dependent. This distinction is important in the context of well-being prediction, because a future application would likely require explanations for individual users, not only global insights about the model.

Therefore, the results suggest that black-box models can be explained to some extent using SHAP and LIME, but that these explanations should not be interpreted as perfectly equivalent to the transparency of a linear model. Ridge Regression remains easier to understand because its coefficients are directly linked to the model structure. In contrast, the explanations of CatBoost, LightGBM, and Stacking are post-hoc approximations of more complex decision mechanisms. They are useful and relatively coherent, but they require more caution, especially when used to interpret individual predictions.

In this sense, the black-box models appear explainable in the context of well-being prediction, but not in the same way as the linear baseline. Their explanations are informative because they reveal consistent patterns of feature importance, particularly at the global level. However, their interpretability remains dependent on external XAI methods, and the local explanations show that the interpretation of a specific prediction can vary depending on the model and the explanation method used.

4.3 Local explanation reliability: faithfulness and robustness

After comparing the explanations produced by the selected models, the analysis focused on two additional dimensions of explanation reliability: faithfulness and robustness. These dimensions are important because an explanation should not only be understandable, but should also reflect the actual behavior of the model and remain stable when the input data are slightly modified.

4.3.1 Faithfulness of local explanations

The faithfulness analysis was based on the modification of *nb friends dispo*, a feature identified as highly influential in several local explanations, especially for the black-box models. The results show that modifying this variable led to a substantial change in the predicted well-being score for CatBoost, LightGBM, and the Stacking Regressor. More specifically, the predicted score increased by 6.469 points for CatBoost, 7.220 points for LightGBM, and 4.053 points for the Stacking Regressor. This indicates that the feature identified as important by the explanation methods also had a meaningful effect on the model output when it was modified.

This result supports the idea that the local explanations of the black-box models are, at least in this specific case, coherent with the actual predictive behavior of the models. If a feature is presented as important in the explanation, then changing its value should affect the prediction. The observed changes in prediction therefore provide a local indication of faithfulness for CatBoost, LightGBM, and Stacking.

The Ridge Regressor shows a different pattern. After modifying *nb friends dispo*, its prediction changed only slightly, decreasing by 0.845 points. This result is coherent with the explanations obtained for Ridge, where *nb friends dispo* was not among the most influential features. Therefore, the small variation observed for Ridge does not necessarily indicate a lack of faithfulness. Rather, it confirms that this model does not rely on this variable in the same way as the black-box models.

However, this faithfulness analysis must be interpreted carefully. It was conducted on a single selected individual and with the modification of only one feature. It therefore provides a local indication of faithfulness, but it does not prove that the explanations are faithful for all individuals or for all important variables. A more complete evaluation would require repeating this perturbation analysis across multiple observations and several influential features.

4.3.2 Robustness of local explanations

The robustness analysis examined whether the local explanations remained stable after the modification of *nb friends dispo*. The results show a high level of stability for both SHAP and LIME explanations. For SHAP, the top-5 most important features remained identical for all four models after the perturbation. The top-10 overlap was also high, ranging from 8/10 to 10/10, and the Spearman rank correlations remained above 0.9 for all models.

The LIME explanations appeared even more stable in this analysis. For all four models, the top-5 and top-10 overlaps were complete, meaning that the same features remained among the most influential variables before and after the modification. The Spearman rank correlations were also almost equal to 1, indicating that the complete feature rankings changed very little.

These results suggest that the local explanations are robust to this specific perturbation. In other words, slightly modifying the selected individual did not substantially change the explanation structure. This is an important result, because unstable explanations would be difficult to trust in a well-being prediction context. If small changes in the input data produced completely different explanations, the practical usefulness of SHAP or LIME would be limited.

Nevertheless, the high robustness observed here should not be overgeneralized. The analysis remains limited to one individual and one modified feature. Moreover, the perturbation was not designed to evaluate all possible forms of instability, but only to assess whether the explanations remained similar after a specific local modification. Therefore, the results suggest good local robustness in the case studied, but they do not establish that SHAP and LIME are robust across the entire dataset.

Overall, the faithfulness and robustness analyses provide encouraging evidence regarding the reliability of the explanations. The feature identified as important by the black-box explanations produced a meaningful change in the predicted score when modified, and the explanations remained stable after this perturbation. However, these findings should be interpreted as local evidence rather than as a general validation of the explanation methods.

4.4 Implications for well-being prediction

The results of this thesis have several implications for the use of machine learning models in well-being prediction. First, they show that well-being can be partially predicted from a reduced set of lifestyle, social, health-related, and contextual variables. The trained models clearly outperform the naive baselines, which indicates that the selected features contain meaningful information for estimating the well-being score. However, the remaining prediction error also shows that well-being is not easy to predict with high precision. This is consistent with the nature of well-being itself, which is subjective, multidimensional, and influenced by factors that may not be fully captured by a short questionnaire.

A second implication concerns the choice of model. The results do not suggest that black-box models should automatically replace simpler models. Although the Stacking Regressor achieved the best average predictive performance, its advantage over CatBoost, LightGBM, and Ridge Regression remained limited. In particular, the Ridge Regressor remained highly competitive while offering intrinsic interpretability. This means that, in a real-world well-being application, the choice between a linear model and a black-box model should not be based only on a small difference in MAE or MSE. It should also take into account the interpretability requirements of the context in which the model will be used.

This is particularly important because well-being prediction belongs to a sensitive domain. A predicted score may influence how users understand their own situation or how professionals interpret an individual profile. Therefore, a model should not only provide accurate predictions; it should also provide explanations that are understandable and reliable. In this respect, the results show that black-box models can be accompanied by meaningful explanations, especially through SHAP and LIME. These methods make it possible to identify which variables contribute most strongly to the prediction and to compare the behavior of complex models with that of a linear baseline.

However, the use of XAI methods does not make black-box models fully transparent. The explanations produced by SHAP and LIME remain post-hoc explanations. For example, in the local SHAP analysis conducted on the selected individual, the feature `nb_friends_dispo` was identified as the most influential feature for the Stacking Regressor, CatBoost, and LightGBM (see Appendix Table 18). Its SHAP contribution was negative for these models, indicating that the fact that the individual reported no immediately available friend contributed to lowering the predicted well-being score. This concrete explanation illustrates how SHAP can indicate not only which feature is important for a specific prediction, but also the direction of its contribution. Similarly, the LIME analysis also ranked `nb_friends_dispo` as the most influential feature for the same black-box models (see Appendix Table 20), confirming that this variable played a central role in the local explanation of this prediction.
They help approximate and interpret model behavior, but they are not equivalent to the direct interpretability of a linear model. This distinction is important for practical use: explanations should be presented as support for understanding the prediction, not as definitive causal interpretations of well-being. For example, if a feature is identified as important, this means that it contributed to the model prediction, not that it necessarily causes a higher or lower level of well-being.

From a practical perspective, the results suggest that different model choices may be appropriate depending on the objective. If the priority is transparency and ease of communication, Ridge Regression remains a strong candidate because its performance is close to that of the best models and its coefficients are directly interpretable. If the priority is to obtain the lowest possible prediction error, then black-box models such as Stacking, CatBoost, or LightGBM may be considered, provided that their predictions are systematically accompanied by XAI analyses. In this case, SHAP and LIME can help make the predictions more understandable and support the critical evaluation of the model.

Overall, the findings suggest that black-box models can be useful for well-being prediction, but only under specific conditions. Their use appears justified when the small gain in performance is considered valuable and when explanation methods are used to examine how the predictions are produced. In this sense, the contribution of XAI is not to completely remove the opacity of black-box models, but to make their behavior more accessible, comparable, and critically interpretable.

4.5 Managerial implications

Beyond its methodological contribution, this thesis also has implications for the field of management. As artificial intelligence systems are increasingly integrated into organizational decision-making processes, the question is no longer only whether a model performs well, but also whether its outputs can be understood, justified, and responsibly used by decision-makers. This is particularly relevant for business engineering, where technical tools must often be evaluated not only in terms of predictive performance, but also in terms of their practical value, organizational acceptability, and governance implications.

The results of this thesis show that black-box models such as Stacking, CatBoost, and LightGBM can slightly improve predictive performance compared to a more transparent linear model. However, this im-

provement remains limited, and the Ridge Regressor remains highly competitive while offering intrinsic interpretability. From a managerial perspective, this finding is important because it shows that the most complex model is not automatically the most appropriate one for implementation. In an organizational context, the choice of a predictive model should therefore depend on a trade-off between accuracy, interpretability, reliability, and ease of communication.

This trade-off is especially important in sensitive domains such as health and well-being. A predicted well-being score may influence how users, professionals, or organizations interpret an individual situation. Therefore, managers and decision-makers cannot rely only on numerical performance indicators such as MAE or MSE. They must also be able to understand which variables influence the prediction, whether these explanations are coherent, and whether the model behaves in a reliable way. In this sense, explainability becomes a condition for the responsible adoption of AI-based decision-support tools.

The explainability results obtained in this thesis are therefore relevant for managerial decision-making. At the global level, SHAP showed that the black-box models relied on similar groups of variables, such as sleep, social relationships, self-management, volunteering, and neighborhood context. This can help decision-makers understand the general logic of the model and assess whether it is aligned with domain knowledge. At the local level, SHAP and LIME made it possible to explain the prediction of a specific individual, although the explanations were more model-dependent. This distinction is important for organizations because global explanations may support model auditing and strategic understanding, while local explanations may support case-by-case interpretation.

However, the results also show that explainability should not be treated as a simple technical add-on. SHAP and LIME provide useful insights, but they do not make black-box models fully transparent. Their explanations remain post-hoc approximations and must therefore be interpreted with caution. For managers, this means that explainability tools should be used as support for critical evaluation, not as automatic justification of model decisions. In practice, the deployment of such models would require clear communication about what the explanations mean, what their limitations are, and how they should be used by non-technical stakeholders.

These findings also have implications for AI governance. In organizations, the implementation of predictive models requires accountability, transparency, and the ability to justify decisions. The results of this thesis suggest that explainability can contribute to these requirements by making model behavior more accessible and by helping identify whether the model relies on coherent patterns. However, responsible implementation would also require additional safeguards, such as external validation, monitoring of model performance over time, fairness analysis, and involvement of domain experts.

Overall, the managerial implication of this thesis is that the adoption of black-box models should be conditional rather than automatic. If the objective is to maximize predictive performance, black-box models may be relevant, but only if their predictions are accompanied by robust explainability analyses. If the priority is transparency, communication, and ease of implementation, a simpler model such as Ridge Regression may remain preferable. Therefore, this thesis contributes to management by showing how explainability can support more informed and responsible choices when organizations consider implementing AI-based decision-support systems.

4.6 Limitations and future research

Several limitations must be considered when interpreting the results of this thesis. First, the predictive comparison was based on five outer cross-validation folds from a single dataset. Although this procedure provides a structured estimate of model performance, it offers only five fold-level values for statistical testing. This is a very limited number of observations and reduces the statistical power of the Friedman, Nemenyi, and Wilcoxon tests. Moreover, these folds should not be interpreted in the same way as several independent datasets. This partly explains why the Stacking Regressor obtained the best average performance but was not significantly better than the closest competing models. Future work could strengthen the statistical comparison by using repeated cross-validation, bootstrapping procedures, or external validation on an independent dataset.

A second limitation concerns the scope of the model search. The analysis focused on a selected set

of models and hyperparameter configurations considered relevant for the research question. However, it was not possible to test all existing machine learning models or all possible combinations of hyperparameters. Other model families, more extensive tuning procedures, or alternative ensemble strategies could potentially lead to different results. Future research could therefore explore additional models and more systematic optimization strategies.

The explainability analysis also presents important limitations. The global SHAP analysis provides useful information about the overall behavior of the selected models, but the local SHAP and LIME analyses were conducted on only one selected individual. As a result, the local explanations cannot be generalized to all individuals in the dataset. The fact that the models showed partial agreement for this observation does not necessarily imply that the same level of agreement would be observed for other profiles. Future research should therefore conduct local explanation analyses on several individuals in order to examine whether the observed patterns remain consistent across different cases. For example, individuals with low, medium, and high predicted well-being scores could be selected and compared.

Similarly, the faithfulness and robustness analyses were limited to one perturbation of one feature, *nb friends dispo*. This choice was relevant because this feature appeared as highly influential for several models, but it does not provide a complete evaluation of explanation reliability. A more comprehensive approach would involve perturbing several important and less important features across multiple individuals, and then measuring whether the resulting prediction changes are consistent with the explanations. This would provide stronger evidence regarding the faithfulness and robustness of SHAP and LIME in this context.

Another limitation is that this thesis focused only on selected dimensions of XAI evaluation. As discussed in the literature review, explainability can be evaluated through many other criteria, such as fairness, complexity, localisation, randomisation, human-grounded evaluation, and application-grounded evaluation. These dimensions were not tested in the present work because they require additional methodological conditions. For example, fairness analyses would require the definition and comparison of relevant subgroups, while human-grounded or application-grounded evaluations would require participants, domain experts, or real users interacting with the explanations. Future research could therefore extend this work by evaluating whether the explanations are not only stable and faithful, but also understandable, useful, and appropriate for the intended audience.

Finally, the results should not be interpreted causally. SHAP and LIME identify features that contribute to model predictions, but they do not prove that these variables cause changes in well-being. This distinction is particularly important in a sensitive domain such as mental health and well-being. For example, if a feature receives a high importance score, this means that the model uses this feature to make predictions, not that changing this feature would necessarily improve an individual's well-being. Future research could therefore combine predictive modeling with longitudinal data, causal inference approaches, or expert validation in order to better distinguish predictive associations from causal mechanisms.

Overall, these limitations do not invalidate the results, but they define the scope within which they should be interpreted. The findings suggest that black-box models can be made partially explainable in the context of well-being prediction, but this explainability remains dependent on the selected models, methods, metrics, individuals, and perturbations. Future work should therefore validate these conclusions on a broader set of observations, conduct local analyses on multiple individuals, apply more systematic perturbation tests, and evaluate additional dimensions of explainability under appropriate methodological conditions.

Conclusion

This thesis examined whether black-box machine learning models can be made explainable in the context of well-being prediction, and how their explainability compares with that of a more transparent linear model. The objective was therefore not only to identify the model with the best predictive performance, but also to assess whether the use of more complex models can be justified when interpretability, reliability, and practical implementation are taken into account.

The results show that black-box models can slightly improve the prediction of well-being. The Stacking Regressor obtained the lowest average prediction error, followed closely by CatBoost, LightGBM, and Ridge Regression. However, the statistical tests did not provide strong evidence that the Stacking Regressor was significantly better than the strongest competing models. This means that, although complex models can perform well, their predictive advantage over the linear baseline remains limited. In particular, Ridge Regression remained highly competitive while offering intrinsic interpretability.

The explainability analysis provides a more nuanced answer to the research question. At the global level, SHAP showed that the selected black-box models relied on highly similar variables, including sleep, social relationships, self-management, volunteering, and neighborhood context. These variables are coherent with the multidimensional nature of well-being and suggest that the models do not rely on arbitrary or incoherent patterns. Compared with the black-box models, Ridge Regression shared several important predictors, but ranked them differently, suggesting that non-linear models may capture additional relationships or interactions in the data.

At the local level, the explanations were more dependent on the model and on the explanation method used. SHAP and LIME showed partial agreement between the black-box models for the selected individual, while Ridge Regression appeared more distinct. The faithfulness analysis showed that modifying an important feature, namely `nb_friends_dispo`, produced a meaningful change in the predictions of the black-box models. The robustness analysis also indicated that the local explanations remained stable after this perturbation. These results suggest that SHAP and LIME can provide useful insights into black-box predictions, but they also confirm that these explanations remain post-hoc approximations rather than direct access to the internal logic of the models.

The research question can therefore be answered in a balanced way. Black-box models are explainable to some extent in the context of well-being prediction, especially when post-hoc methods such as SHAP and LIME are used to analyze their behavior. However, this explainability remains partial, method-dependent, and less direct than the transparency provided by a linear model. As a result, the use of black-box models cannot be justified only by a small improvement in predictive performance. It must also be justified by the quality, stability, and usefulness of the explanations that accompany the predictions.

From a managerial perspective, this conclusion is particularly important. In organizations, the adoption of artificial intelligence systems does not depend only on technical performance. It also depends on whether decision-makers can understand, communicate, and justify the outputs produced by these systems. This is especially relevant in sensitive domains such as health and well-being, where a predicted score may influence how users, professionals, or organizations interpret an individual situation. In such a context, explainability becomes a condition for responsible implementation rather than a secondary technical feature.

The findings of this thesis therefore suggest that managers should not automatically favor the most complex model. If transparency, communication, and ease of implementation are priorities, a simpler model such as Ridge Regression may remain preferable, especially given its competitive performance. If the objective is to minimize prediction error, black-box models such as Stacking, CatBoost, or LightGBM may be considered, but only if their predictions are systematically accompanied by explainability analyses. In this sense, the managerial value of XAI lies in its ability to support informed decision-making, improve trust, and make algorithmic systems more accountable in practice.

This thesis also highlights the importance of AI governance. The deployment of predictive models in organizations requires more than model selection. It requires clear procedures for validating model be-

havior, monitoring performance over time, communicating limitations to stakeholders, and ensuring that predictions are not interpreted as causal conclusions. SHAP and LIME can contribute to this process by making model behavior more accessible, but they cannot replace human judgment, domain expertise, or organizational responsibility.

Several limitations must nevertheless be acknowledged. The predictive comparison was based on five cross-validation folds from a single dataset, which limits the strength of the statistical conclusions. The local explainability, faithfulness, and robustness analyses were conducted on one selected individual and one feature perturbation, which means that the findings cannot be generalized to all possible profiles. Future research should therefore extend the local analyses to several individuals, test faithfulness and robustness across multiple features, and include additional dimensions of XAI evaluation, such as fairness, human-grounded evaluation, and expert validation.

Overall, this work shows that black-box models can be useful for well-being prediction, but only under specific conditions. Their use should be based on a careful balance between predictive performance, interpretability, reliability, and responsible application. The main contribution of this thesis is therefore not to show that black-box models are always preferable, but to clarify when their use may be acceptable and how explainability methods can support their critical and managerial evaluation in a sensitive real-world context.

Bibliography

[1] Chirag Agarwal et al. "OpenXAI: Towards a Transparent Evaluation of Post hoc Model Explanations". In: *Advances in Neural Information Processing Systems*. Datasets and Benchmarks Track. 2024.

[2] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[3] Haewon Byeon. "Exploring Factors for Predicting Anxiety Disorders of the Elderly Living Alone in South Korea Using Interpretable Machine Learning: A Population-Based Study". In: *International Journal of Environmental Research and Public Health* 18.14 (2021), p. 7625. DOI: 10.3390/ijerph18147625.

[4] Dulce Canha et al. "A Functionally-Grounded Benchmark Framework for XAI Methods: Insights and Foundations from a Systematic Literature Review". In: *ACM Computing Surveys* 57.12 (2025), Article 320. DOI: 10.1145/3737445.

[5] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[6] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. "The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation". In: *PeerJ Computer Science* 7 (2021), e623. DOI: 10.7717/peerj-cs.623.

[7] Finale Doshi-Velez and Been Kim. "Towards a Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[8] Raghavendra Dwivedi et al. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions". In: *ACM Computing Surveys* 55.9 (2023), p. 194. DOI: 10.1145/3561048.

[9] Riccardo Guidotti et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (2018), p. 93. DOI: 10.1145/3236009.

[10] Anna Hedström et al. "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond". In: *Journal of Machine Learning Research* 24.34 (2023), pp. 1–11.

[11] Tim Hulsen. "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare". In: *AI* 4.3 (2023), pp. 652–666. DOI: 10.3390/ai4030034.

[12] Dana Jaber et al. "Medically-Oriented Design for Explainable AI for Stress Prediction from Physiological Measurements". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), p. 38. DOI: 10.1186/s12911-022-01772-2.

[13] Ishan P. Jha et al. "Learning the Mental Health Impact of COVID-19 in the United States with Explainable Artificial Intelligence: Observational Study". In: *JMIR Mental Health* 8.4 (2021), e25097. DOI: 10.2196/25097.

[14] Daniel W. Joyce et al. "Explainable Artificial Intelligence for Mental Health through Transparency and Interpretability for Understandability". In: *npj Digital Medicine* 6.1 (2023), p. 6. DOI: 10.1038/s41746-023-00751-9.

[15] Isha Kaur et al. "Enhancing Explainability in Predicting Mental Health Disorders Using Human–Machine Interaction". In: *Multimedia Tools and Applications* (2024). Advance online publication. DOI: 10.1007/s11042-024-18346-1.

[16] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 3146–3154.

[17] Corey L. M. Keyes. *Brief Description of the Mental Health Continuum Short Form (MHC-SF)*. Tech. rep. Technical report. Emory University, 2009. URL: http://www.sociology.emory.edu/ckeyes/.

[18] Taehoon Kim et al. "Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 2022, pp. 1–20. DOI: 10.1145/3491102.3517701.

[19] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[20] Chrysanthi Ntakolia et al. "An Explainable Machine Learning Approach for COVID-19's Impact on Mood States of Children and Adolescents during the First Lockdown in Greece". In: *Healthcare* 10.1 (2022), p. 149. DOI: 10.3390/healthcare10010149.

[21] Eva Paraschou, Sofia Yfantidou, and Athena Vakali. "UnStressMe: Explainable Stress Analytics and Self-tracking Data Visualizations". In: *Proceedings of the 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*. 2023.

[22] Venkata Pendyala and Hyeonwoo Kim. "Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI". In: *Electronics* 13.6 (2024), p. 1025. DOI: 10.3390/electronics13061025.

[23] Marius-Constantin Popescu et al. "Multilayer Perceptron and Neural Networks". In: *WSEAS Transactions on Circuits and Systems* 8.7 (2009), pp. 579–588.

[24] Liudmila Prokhorenkova et al. "CatBoost: Unbiased Boosting with Categorical Features". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 6639–6649.

[25] Hassan Ramchoun et al. "Multilayer Perceptron: Architecture Optimization and Training". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 4.1 (2016), pp. 26–30. DOI: 10.9781/ijimai.2016.415.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[27] Alex J. Smola and Bernhard Schölkopf. "A Tutorial on Support Vector Regression". In: *Statistics and Computing* 14.3 (2004), pp. 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

[28] Kai Ming Ting and Ian H. Witten. *Stacked Generalization: When Does It Work?* Working Paper 97/3. Hamilton, New Zealand: Department of Computer Science, University of Waikato, 1997.

[29] Bogdan Trawinski et al. "Nonparametric Statistical Analysis for Multiple Comparison of Machine Learning Regression Algorithms". In: *International Journal of Applied Mathematics and Computer Science* 22.4 (2012), pp. 867–881. DOI: 10.2478/v10006-012-0064-z.

[30] Flore Vancompernolle Vromman et al. "Explainable AI for Well-Being Prediction From Lifestyle Data: 2-Study Design". In: *JMIR Mental Health* 13 (2026), e88750. DOI: 10.2196/88750.

[31] Robert O. Weber et al. "XAI Is in Trouble". In: *AI Magazine* 45.4 (2024), pp. 300–316. DOI: 10.1002/aaai.12184.

[32] Xingyu Zhao et al. "BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021)*. Vol. 161. Proceedings of Machine Learning Research. PMLR, 2021, pp. 887–896.

Appendix

Description of the selected features

Table 1: Description of the selected features used in the XGBoost model. Some original questionnaire items were transformed into several encoded features, especially when categorical or multiple-choice variables were converted into binary indicator variables.

Selected feature	Original questionnaire item	Answer or coding represented
nb_friends_dispo	Please indicate the number of immediately available friends with whom you can talk frankly, in person, by phone, or by text, without having to watch what you say.	Number of immediately available friends.
autogestion_9	Here is a list of ways you may use to feel well or better, maintain good mental health, or avoid psychological difficulties. For each tool, indicate how often you have used it in the past month: "I have a healthy diet."	Never = 1; Rarely = 2; Often = 3; Very often = 4.
maladies_20	Long-term health conditions diagnosed by a health professional: chronic renal failure.	Binary indicator: Chronic renal failure = 1.
travail_domaine_6	Which one of the following categories best describes your field of employment?	Binary indicator for "Arts, Culture, Sport and Recreation".
sommeil_1	During the past seven days, how would you rate your overall sleep quality?	Sleep quality score.
maladies_19	Long-term health conditions diagnosed by a health professional: chronic fatigue syndrome.	Binary indicator: Chronic fatigue syndrome = 1.
issue_ai_data_3_4.0	Agreement with the statement: "I agree that the government uses my numeric personal data, if it is for the public good."	Binary indicator for the response category "Strongly agree".
origines_ethniques_2.0	Which of the following categories best describes you?	Binary indicator for the response category "Black".
quartier_domicile_3	Perception of the neighborhood where the respondent lives: "The neighborhood's population is friendly."	Strongly disagree = 1; Somewhat disagree = 2; Somewhat agree = 3; Strongly agree = 4.
SoutSup_6	Please answer the following statement by selecting the one that applies to you: "There are times to discuss the difficulties involved in carrying out our task."	Strongly agree = 1; Agree = 2; More or less in agreement = 3; Disagree = 4; Strongly disagree = 5.
maladies_21	Long-term health conditions diagnosed by a health professional: liver disease or gallbladder problems.	Binary indicator: Liver disease or gallbladder problems = 1.
style_3.0	What is your clothing style?	Binary indicator for the response category "Classical".
travail_domaine_10	Which one of the following categories best describes your field of employment?	Binary indicator for "Manufacturing and utilities".
travail_domaine_1	Which one of the following categories best describes your field of employment?	Binary indicator for "Management".

Continued on next page

Selected feature	Original questionnaire item	Answer or coding represented
travail_domaine_3	Which one of the following categories best describes your field of employment?	Binary indicator for "Natural and applied sciences and related fields".
consult_who_3	Whom did you see or talk to regarding your emotional or mental health?	Binary indicator for "Psychologist".
consult_who_6	Whom did you see or talk to regarding your emotional or mental health?	Binary indicator for "Other".
married_5.0	What is your marital status?	Binary indicator for "Divorced/separated".
act_volunteer	How often do you volunteer or involve yourself in a cause?	Never = 1; Almost never = 2; Sometimes = 3; Often = 4; Very often = 5.
act_friends	How often do you do activities with one or more friend(s)?	Never = 1; Almost never = 2; Sometimes = 3; Often = 4; Very often = 5.
quartier_opportunite	My neighborhood offers many opportunities, such as infrastructure, sports and social activities, services or shops, to take care of my health.	Strongly disagree = 1; Disagree = 2; Agree = 3; Strongly agree = 4.
consult_who_5	Whom did you see or talk to regarding your emotional or mental health?	Binary indicator for "Social worker or counselor".
maladies_16	Long-term health conditions diagnosed by a health professional: cancer.	Binary indicator: Cancer = 1.
chronotype_3.0	Self-assessment of chronotype by choosing a graph representing the evolution of alertness over the course of the day.	Binary indicator for the response category "Highly active".
travail_domaine_2	Which one of the following categories best describes your field of employment?	Binary indicator for "Business, finance and administration".
origines_ethniques_4.0	Which of the following categories best describes you?	Binary indicator for the response category "Asian".
style_8.0	What is your clothing style?	Binary indicator for the response category "Sporty".
maladies_22	Long-term health conditions diagnosed by a health professional: stomach or intestinal ulcers.	Binary indicator: Stomach or intestinal ulcers = 1.
chronotype_2.0	Self-assessment of chronotype by choosing a graph representing the evolution of alertness over the course of the day.	Binary indicator for the response category "Evening".
act_nature_1	How often do you spend time in green or natural environments? May to September.	Never = 1; Almost never = 2; Sometimes = 3; Often = 4; Very often = 5.
car_model_4.0	Which of the following car models do you happen to use most often?	Binary indicator for "Pickup".
smoking	How often do you smoke cigarettes and/or vape?	Never = 1; A few times a year = 2; Once a month = 3; Once a week = 4; A few times a week = 5; Once a day = 6; More than once a day = 7.

Continued on next page

Selected feature	Original questionnaire item	Answer or coding represented
LatDec_3	In the same context, please respond to the following statement: "I can decide when to take a break."	Strongly agree = 1; Agree = 2; More or less in agreement = 3; Disagree = 4; Strongly disagree = 5.
style_2.0	What is your clothing style?	Binary indicator for the response category "Elegant".
travail_domaine_8	Which one of the following categories best describes your field of employment?	Binary indicator for "Trades, transport, equipment operators and related occupations".
maladies_15	Long-term health conditions diagnosed by a health professional: diabetes.	Binary indicator: Diabetes = 1.

Detailed predictive results by model

This appendix reports the detailed fold-by-fold predictive results obtained for each model. For each model, the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and, when applicable, the selected preprocessing steps and hyperparameters are reported across the five cross-validation folds.

Baseline models

Table 2: Mean Regressor performance across the five folds

Fold	MAE	MSE
1	18.5520	502.3985
2	16.8971	430.4231
3	18.4479	499.4209
4	18.1730	484.3616
5	17.0881	454.6369
Mean	**17.8316**	**474.2482**

Table 3: Random Regressor performance across the five folds

Fold	MAE	MSE
1	32.1466	1467.1014
2	31.0180	1376.7310
3	30.5458	1376.1947
4	30.4683	1397.6234
5	30.1428	1322.1106
Mean	**30.8643**	**1387.9522**

Detailed results for machine learning models

Table 4: XGBoost Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Depth / Subsample
1	14.6095	335.2969	SimpleImputer	StandardScaler	3 / 0.8
2	13.9877	313.3115	SimpleImputer	MinMaxScaler	3 / 0.8
3	15.0843	353.6073	SimpleImputer	StandardScaler	3 / 0.8
4	14.9531	339.7564	KNNImputer	MinMaxScaler	3 / 0.8
5	14.9619	364.4686	KNNImputer	MinMaxScaler	3 / 0.8
Mean	**14.7193**	**341.2881**	–	–	–

Table 5: LightGBM Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Learning rate	Estimators / Leaves
1	14.8531	333.3220	SimpleImputer	MinMaxScaler	0.01	200 / 31
2	13.7020	289.8182	SimpleImputer	StandardScaler	0.10	50 / 31
3	14.5231	331.4639	SimpleImputer	MinMaxScaler	0.10	50 / 50
4	14.0990	296.0140	SimpleImputer	StandardScaler	0.10	50 / 31
5	14.1353	305.9530	SimpleImputer	MinMaxScaler	0.01	200 / 50
Mean	**14.2625**	**313.2816**	–	–	–	–

Table 6: Support Vector Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Kernel	C / Epsilon
1	14.9052	338.2210	KNNImputer	MinMaxScaler	Linear	1 / 0.2
2	14.4084	322.5085	SimpleImputer	MinMaxScaler	RBF	10 / 0.5
3	14.1775	324.1989	SimpleImputer	MinMaxScaler	RBF	10 / 0.1
4	14.5395	318.4004	SimpleImputer	MinMaxScaler	RBF	10 / 0.5
5	14.2648	305.6379	KNNImputer	MinMaxScaler	Linear	1 / 0.1
Mean	**14.4591**	**321.7933**	–	–	–	–

Table 7: MLP Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Hidden layer	Solver
1	15.3575	366.5990	KNNImputer	StandardScaler	25	Adam
2	15.1828	361.3835	SimpleImputer	StandardScaler	25	Adam
3	14.1775	324.1989	SimpleImputer	StandardScaler	25	Adam
4	16.1870	397.1833	KNNImputer	StandardScaler	25	Adam
5	*15.8075*	380.9989	SimpleImputer	StandardScaler	25	Adam
Mean	***15,3425***	**366.0727**	–	–	–	–

Table 8: Random Forest Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Estimators	Depth
1	15.0514	343.2372	SimpleImputer	MinMaxScaler	100	10
2	14.0772	311.7907	SimpleImputer	StandardScaler	100	10
3	15.0323	346.4516	SimpleImputer	StandardScaler	100	10
4	14.9942	334.1328	SimpleImputer	StandardScaler	50	10
5	14.3722	313.8298	KNNImputer	MinMaxScaler	100	10
Mean	**14.7055**	**329.8884**	–	–	–	–

Table 9: KNN Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Neighbors	Weights / p
1	15.3827	362.5754	SimpleImputer	StandardScaler	7	Uniform / 1
2	15.4652	359.3565	KNNImputer	StandardScaler	7	Uniform / 1
3	15.7302	392.0724	SimpleImputer	MinMaxScaler	7	Uniform / 1
4	15.9742	392.3957	KNNImputer	StandardScaler	7	Uniform / 1
5	15.1512	348.1965	KNNImputer	StandardScaler	7	Uniform / 1
Mean	**15.5407**	**370.9193**	–	–	–	–

Table 10: ElasticNet Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Alpha	L1 ratio
1	14.8680	339.7732	SimpleImputer	StandardScaler	0.1	0.2
2	14.2976	312.2447	KNNImputer	StandardScaler	0.1	0.2
3	17.5640	448.0892	KNNImputer	MinMaxScaler	1.0	0.8
4	14.5305	317.6987	KNNImputer	MinMaxScaler	0.1	0.8
5	14.3521	308.8275	SimpleImputer	StandardScaler	0.1	0.2
Mean	**15.1224**	**345.3267**	–	–	–	–

Table 11: CatBoost Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Depth	Iterations	Learning rate
1	14.1610	307.5779	SimpleImputer	MinMaxScaler	4	100	0.1
2	13.7003	288.2281	KNNImputer	MinMaxScaler	4	100	0.1
3	14.2166	308.6706	KNNImputer	StandardScaler	4	100	0.1
4	13.9823	296.0140	SimpleImputer	StandardScaler	4	100	0.1
5	14.1214	305.9530	SimpleImputer	StandardScaler	4	100	0.1
Mean	**14.0363**	**301.2887**	–	–	–	–	–

Table 12: Stacking Regressor performance and selected hyperparameters across the five folds

Fold	MAE	MSE	Imputer	Scaler	Base estimators	Final estimator
1	14.2118	314.9536	KNNImputer	MinMaxScaler	CatBoost + Ridge	Ridge
2	13.6202	287.3347	SimpleImputer	MinMaxScaler	CatBoost + Ridge	Ridge
3	14.1485	310.9719	KNNImputer	MinMaxScaler	CatBoost + Ridge	CatBoost
4	13.8617	290.7375	SimpleImputer	MinMaxScaler	CatBoost + Ridge	Ridge
5	13.9651	296.3279	SimpleImputer	MinMaxScaler	CatBoost + Ridge	Ridge
Mean	**13.9615**	**300.0651**	–	–	–	–

Detailed statistical test results

This appendix presents additional statistical results used to compare the predictive performance of the regression models. These tables provide complementary details to the main statistical analysis reported in the results section.

Fold-level model ranks

Table 13: Fold-level ranks of regression models based on MAE.

Fold	Stacking	CatBoost	LightGBM	Ridge	SVR	XGBoost	ElasticNet	Random Forest	MLP	KNN	Mean	Random
Fold 1	2.0	1.0	5.0	4.0	6.0	3.0	7.0	8.0	9.0	10.0	11.0	12.0
Fold 2	1.0	2.0	3.0	5.0	7.0	4.0	6.0	8.0	9.0	10.0	11.0	12.0
Fold 3	1.0	4.0	6.0	3.0	2.0	9.0	12.0	8.0	2.0	10.0	11.0	12.0
Fold 4	1.0	2.0	3.0	4.0	6.0	8.0	5.0	9.0	12.0	10.0	11.0	12.0
Fold 5	1.0	2.0	3.0	7.0	5.0	8.0	6.0	9.0	10.0	11.0	12.0	12.0

Note. This table reports the rank obtained by each regression model in each outer cross-validation fold, based on MAE values. Lower ranks indicate better predictive performance. A rank of 1 corresponds to the model with the lowest MAE in a given fold. The model names were shortened for readability: SVR refers to Support Vector Regressor.

Pairwise Nemenyi post-hoc comparisons

Table 14: Pairwise Nemenyi post-hoc comparisons between regression models.

	Stacking	CatBoost	LightGBM	Ridge	SVR	XGBoost	ElasticNet	Random Forest	MLP	KNN	Mean	Random
Stacking	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	**	***	***
CatBoost	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	*	***	***
LightGBM	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	*	**
Ridge	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	*
SVR	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
XGBoost	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
ElasticNet	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Random Forest	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
MLP	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
KNN	**	*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mean	***	***	*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Random	***	***	**	*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Note. This table reports the pairwise Nemenyi post-hoc comparisons between regression models based on MAE ranks. *** indicates $p < 0.01$, ** indicates $p < 0.05$, * indicates $p < 0.1$, and n.s. indicates a non-significant difference.

Pairwise Wilcoxon signed-rank comparisons

Table 15: Pairwise Wilcoxon signed-rank comparisons between regression models with Holm correction.

	Stacking	CatBoost	LightGBM	Ridge	SVR	XGBoost	ElasticNet	Random Forest	MLP	KNN	Mean	Random
Stacking	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
CatBoost	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
LightGBM	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Ridge	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SVR	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
XGBoost	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
ElasticNet	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.	n.s.
Random Forest	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.	n.s.
MLP	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.	n.s.
KNN	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.	n.s.
Mean	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–	n.s.
Random	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	–

Note. This table reports the pairwise Wilcoxon signed-rank comparisons between regression models based on MAE values. The p-values were corrected using the Holm method to account for multiple comparisons. *** indicates $p < 0.01$, ** indicates $p < 0.05$, * indicates $p < 0.1$, and n.s. indicates a non-significant difference. The symbol – indicates the comparison of a model with itself.

Global SHAP analyse

Table 16: Comparison of feature importance rankings across the four models

Rank	Feature	Stacking	CatBoost	LightGBM	Ridge
1	sommeil_1	1	2	2	1
2	nb_friends_dispo	2	1	1	34
3	autogestion_9	3	3	3	2
4	quartier_domicile_3	4	4	6	3
5	act_volunteer	5	6	5	6
6	SoutSup_6	6	5	4	4
7	act_friends	7	8	7	5
8	act_nature_1	8	7	10	7
9	quartier_opportunite	9	9	8	8
10	issue_ai_data_3_4.0	10	12	12	11
11	smoking	11	10	9	12
12	origines_ethniques_4.0	12	11	13	9
13	LatDec_3	13	13	11	14
14	chronotype_3.0	14	16	16	13
15	maladies_15	15	14	15	10
16	maladies_19	16	23	20	20
17	origines_ethniques_2.0	17	15	14	16
18	consult_who_5	18	22	21	18
19	married_5.0	19	19	22	21
20	style_2.0	20	25	26	25
21	style_3.0	21	17	19	17
22	travail_domaine_8	22	27	23	19
23	travail_domaine_1	23	21	17	22
24	travail_domaine_2	24	24	18	15
25	car_model_4.0	25	18	25	36
26	chronotype_2.0	26	20	27	27
27	travail_domaine_3	27	28	28	29
28	style_8.0	28	26	24	24
29	maladies_21	29	34	34	23
30	maladies_22	30	29	29	28
31	consult_who_6	31	35	33	33
32	maladies_20	32	36	36	30
33	maladies_16	33	30	32	32
34	consult_who_3	34	32	30	26
35	travail_domaine_10	35	31	31	31
36	travail_domaine_6	36	33	35	35

Local SHAP analysis

Table 17: Values of the selected individual used for the local SHAP and LIME explanation analysis. Empty cells are reported as "–".

Variable	Value	Variable	Value
travail_domaine_1	–	travail_domaine_2	–
travail_domaine_3	–	travail_domaine_6	–
travail_domaine_8	–	travail_domaine_10	–
origines_ethniques	7.0	married	1.0
car_model	11.0	smoking	1.0
act_friends	1.0	act_volunteer	3.0
act_nature_1	5.0	style	–
maladies_15	–	maladies_16	–
maladies_19	–	maladies_20	–
maladies_21	–	maladies_22	–
autogestion_9	3.0	sommeil_1	7.0
chronotype	1.0	LatDec_3	–
SoutSup_6	–	quartier_domicile_3	1.0
quartier_opportunite	1.0	consult_who_3	–
consult_who_5	–	consult_who_6	–
nb_friends_dispo	0.0	issue_ai_data_3	1

Table 18: Ranking of local SHAP feature contributions for the selected individual across the four models. Ranks are based on the absolute local SHAP value. A lower rank indicates a stronger contribution to the predicted score.

Feature	Stacking	CatBoost	LightGBM	Ridge
nb_friends_dispo	1	1	1	33
quartier_domicile_3	2	2	2	1
act_friends	3	3	3	2
act_nature_1	4	4	6	5
quartier_opportunite	5	7	8	3
act_volunteer	6	8	5	6
autogestion_9	7	5	4	8
sommeil_1	8	6	14	4
issue_ai_data_3_4.0	9	21	11	12
travail_domaine_1	10	14	15	18
LatDec_3	11	18	16	26
SoutSup_6	11	20	13	7
car_model_4.0	11	30	26	34
chronotype_2.0	11	22	24	24
chronotype_3.0	11	10	21	13
consult_who_3	11	28	31	29
consult_who_5	11	15	18	19
consult_who_6	11	34	32	35
maladies_15	11	19	19	9
maladies_16	11	34	29	30
maladies_19	11	25	22	28
maladies_20	11	34	32	31
maladies_21	11	33	32	21
maladies_22	11	31	30	25
married_5.0	11	23	20	17
origines_ethniques_2.0	11	16	17	16
origines_ethniques_4.0	11	12	9	10
smoking	11	13	10	11
style_2.0	11	9	25	23
style_3.0	11	11	7	15
style_8.0	11	17	27	19
travail_domaine_10	11	24	28	32

Continued on next page

Feature	Stacking	CatBoost	LightGBM	Ridge
travail_domaine_2	11	29	12	14
travail_domaine_3	11	27	32	27
travail_domaine_6	11	32	32	35
travail_domaine_8	11	26	23	22

Table 19: Local SHAP values for the selected individual across the four models. Positive values increase the predicted well-being score, while negative values decrease it.

Feature	CatBoost	LightGBM	Ridge	Stacking
Predicted score	**42.143**	**42.046**	**50.051**	**42.866**
nb_friends_dispo	-13.335	-13.835	0.007	-8.889
quartier_domicile_3	-5.354	-3.811	-7.158	-5.818
act_friends	-2.978	-3.204	-5.789	-3.861
act_nature_1	2.236	1.558	1.880	2.956
quartier_opportunite	-1.703	0.410	-3.061	-1.915
autogestion_9	1.987	2.258	0.687	1.018
act_volunteer	1.480	1.639	1.516	1.196
sommeil_1	1.717	-0.216	2.466	0.645
SoutSup_6	0.115	0.223	1.049	0.000
issue_ai_data_3_4.0	-0.109	-0.264	-0.435	-0.578
style_3.0	0.351	-0.680	0.236	0.000
origines_ethniques_4.0	0.267	0.286	0.635	0.000
smoking	0.248	0.266	0.543	0.000
maladies_15	-0.136	-0.163	-0.648	0.000
chronotype_3.0	-0.491	-0.130	-0.249	0.000
travail_domaine_1	-0.230	-0.188	-0.177	-0.265
style_2.0	0.495	-0.084	-0.116	0.000
origines_ethniques_2.0	-0.169	-0.183	-0.222	0.000
consult_who_5	0.186	0.182	0.174	0.000
travail_domaine_2	0.025	0.230	0.242	0.000
married_5.0	0.100	0.150	0.201	0.000
LatDec_3	-0.138	-0.184	0.092	0.000
style_8.0	-0.145	-0.055	-0.174	0.000
chronotype_2.0	0.104	0.088	0.112	0.000
travail_domaine_8	0.045	0.094	0.142	0.000
maladies_19	0.049	0.123	0.073	0.000
maladies_21	-0.001	0.000	0.170	0.000
travail_domaine_10	-0.069	-0.041	-0.040	0.000
maladies_22	0.013	0.008	0.102	0.000
travail_domaine_3	0.033	0.000	0.076	0.000
consult_who_3	-0.026	0.006	0.056	0.000
maladies_16	0.000	0.038	0.049	0.000
car_model_4.0	-0.015	-0.056	0.006	0.000
maladies_20	0.000	0.000	-0.044	0.000
travail_domaine_6	0.003	0.000	0.000	0.000
consult_who_6	0.000	0.000	-0.000	0.000

LIME analysis

Table 20: Ranking of local LIME feature contributions for the selected individual across the four models. Ranks are based on the absolute local LIME value. A lower rank indicates a stronger contribution to the predicted score.

Feature	Stacking	CatBoost	LightGBM	Ridge
nb_friends_dispo	1	1	1	35
origines_ethniques_2.0	2	2	2	6
issue_ai_data_3_4.0	3	4	6	4
maladies_19	4	8	3	1
quartier_domicile_3	5	3	5	8
style_2.0	6	7	8	5
act_friends	7	6	7	9
consult_who_5	8	10	9	7
travail_domaine_3	9	12	23	3
maladies_15	10	16	12	13
act_nature_1	11	9	11	17
autogestion_9	12	5	4	27
style_8.0	13	17	10	11
maladies_21	14	22	30	2
quartier_opportunite	15	14	14	16
origines_ethniques_4.0	16	13	13	12
travail_domaine_8	17	21	15	10
chronotype_3.0	18	11	19	15
married_5.0	19	18	20	14
style_3.0	20	20	16	22
act_volunteer	21	15	17	26
travail_domaine_2	22	23	18	20
smoking	23	19	22	23
maladies_20	24	35	24	18
maladies_22	25	34	36	19
travail_domaine_10	26	26	26	25
chronotype_2.0	27	28	32	31
consult_who_6	28	25	33	21
travail_domaine_1	29	27	21	28
consult_who_3	30	32	28	24
travail_domaine_6	31	31	31	30
sommeil_1	32	30	34	32
maladies_16	33	36	25	29
LatDec_3	34	29	27	33
SoutSup_6	35	24	29	36
car_model_4.0	36	33	35	34

Table 21: Local LIME values for the selected individual across the four models. Positive values increase the predicted well-being score, while negative values decrease it.

Feature	CatBoost	LightGBM	Ridge	Stacking
Predicted score	**42.143**	**42.046**	**50.051**	**42.866**
nb_friends_dispo	-9.895	-8.649	0.196	-6.649
origines_ethniques_2.0	-4.968	-5.955	-5.984	-5.818
issue_ai_data_3_4.0	-4.599	-4.106	-6.048	-5.369
maladies_19	3.037	5.911	7.296	5.285
quartier_domicile_3	-4.747	-4.733	-5.732	-5.264
style_2.0	3.682	3.822	-6.002	4.581
act_friends	-3.984	-4.075	-4.998	-4.245

Continued on next page

Feature	CatBoost	LightGBM	Ridge	Stacking
consult_who_5	2.661	3.127	5.811	4.094
travail_domaine_3	2.500	-1.154	6.214	3.728
maladies_15	-1.952	-2.534	-4.499	-3.381
act_nature_1	2.691	2.720	3.064	3.320
autogestion_9	4.040	5.067	1.629	3.133
style_8.0	1.901	3.111	-4.794	3.130
maladies_21	1.055	0.411	6.603	3.025
quartier_opportunite	-2.255	-2.507	-3.112	-2.866
origines_ethniques_4.0	2.284	2.529	4.577	2.854
travail_domaine_8	1.216	2.485	4.803	2.801
chronotype_3.0	-2.557	-2.108	-3.159	-2.494
married_5.0	1.750	1.759	3.204	2.358
style_3.0	1.483	2.441	2.610	2.082
act_volunteer	2.119	2.383	1.875	1.999
travail_domaine_2	0.977	2.228	2.821	1.988
smoking	1.501	1.432	2.143	1.872
maladies_20	0.222	1.142	-2.952	-1.281
maladies_22	0.273	0.002	2.882	1.125
travail_domaine_10	-0.671	-1.066	-2.003	-1.006
chronotype_2.0	0.622	0.318	1.065	0.956
consult_who_6	-0.830	-0.314	2.695	0.887
travail_domaine_1	-0.640	-1.727	-1.469	-0.874
consult_who_3	0.413	0.442	2.054	0.859
travail_domaine_6	-0.418	-0.340	-1.170	-0.723
sommeil_1	0.478	-0.222	0.902	0.720
maladies_16	0.102	1.092	1.361	0.633
LatDec_3	-0.572	-0.461	-0.621	-0.414
SoutSup_6	0.922	0.432	0.043	0.405
car_model_4.0	0.343	-0.209	0.396	0.139

Features coefficients with the modified individual

Table 22: Values of the selected individual modified, used for the faithfulness and robustness analysis. Empty cells are reported as "–".

Variable	Value	Variable	Value
travail_domaine_1	–	travail_domaine_2	–
travail_domaine_3	–	travail_domaine_6	–
travail_domaine_8	–	travail_domaine_10	–
origines_ethniques	7.0	married	1.0
car_model	11.0	smoking	1.0
act_friends	1.0	act_volunteer	3.0
act_nature_1	5.0	style	–
maladies_15	–	maladies_16	–
maladies_19	–	maladies_20	–
maladies_21	–	maladies_22	–
autogestion_9	3.0	sommeil_1	7.0
chronotype	1.0	LatDec_3	–
SoutSup_6	–	quartier_domicile_3	1.0
quartier_opportunite	1.0	consult_who_3	–
consult_who_5	–	consult_who_6	–
nb_friends_dispo	1.0	issue_ai_data_3	1

LIME

Table 23: Local LIME values for the modified individual across the four models. Positive values increase the predicted well-being score, while negative values decrease it.

Feature	CatBoost	LightGBM	Ridge	Stacking
Predicted score	48.612	49.266	49.206	46.919
nb_friends_dispo	-9.890	-8.643	0.186	-6.646
origines_ethniques_2.0	-4.963	-5.950	-5.957	-5.815
quartier_domicile_3	-4.743	-4.730	-5.736	-5.262
issue_ai_data_3_4.0	-4.596	-4.103	-6.037	-5.367
autogestion_9	4.043	5.071	1.637	3.134
act_friends	-3.979	-4.069	-5.007	-4.242
style_2.0	3.721	3.866	-5.963	4.606
maladies_19	3.039	5.913	7.276	5.286
act_nature_1	2.698	2.728	3.066	3.324
consult_who_5	2.662	3.128	5.812	4.095
chronotype_3.0	-2.554	-2.105	-3.150	-2.492
travail_domaine_3	2.502	-1.152	6.185	3.729
origines_ethniques_4.0	2.286	2.532	4.563	2.856
quartier_opportunite	-2.250	-2.502	-3.105	-2.863
act_volunteer	2.124	2.389	1.871	2.003
maladies_15	-1.950	-2.532	-4.497	-3.380
style_8.0	1.930	3.144	4.757	3.149
married_5.0	1.751	1.760	3.199	2.359
smoking	1.503	1.434	2.158	1.873
style_3.0	1.498	2.458	-2.628	2.092
travail_domaine_8	1.218	2.487	4.760	2.802
maladies_21	1.056	0.412	6.643	3.026
travail_domaine_2	0.979	2.230	2.819	1.989
SoutSup_6	0.925	0.436	0.054	0.407
consult_who_6	-0.829	-0.314	2.734	0.887
travail_domaine_10	-0.670	-1.064	-1.967	-1.006
travail_domaine_1	-0.640	-1.726	-1.493	-0.874
chronotype_2.0	0.623	0.320	1.068	0.957
LatDec_3	-0.569	-0.457	-0.434	-0.412
sommeil_1	0.481	-0.218	0.903	0.722
travail_domaine_6	-0.416	-0.339	-1.177	-0.722
consult_who_3	0.415	0.443	2.049	0.859
car_model_4.0	0.345	-0.206	0.408	0.140
maladies_22	0.274	0.003	2.928	1.125
maladies_20	0.224	1.144	-3.022	-1.280
maladies_16	0.104	1.095	1.345	0.635

Local SHAP

Table 24: Local SHAP values for the modified individual across the four models. Positive values increase the predicted well-being score, while negative values decrease it.

Feature	CatBoost	LightGBM	Ridge	Stacking
Predicted score	48.612	49.266	49.206	46.919
nb_friends_dispo	-8.647	-7.578	0.007	-5.865
quartier_domicile_3	-5.025	-4.780	-7.158	-5.578
act_friends	-2.925	-2.825	-5.789	-3.773
act_nature_1	2.242	1.314	1.880	2.953
autogestion_9	1.987	2.411	0.687	1.073
sommeil_1	1.973	0.491	2.466	0.788

Continued on next page

Feature	CatBoost	LightGBM	Ridge	Stacking
act_volunteer	1.773	2.077	1.516	1.369
quartier_opportunite	-1.635	0.007	-3.061	-1.658
style_2.0	0.526	-0.067	-0.116	0.000
chronotype_3.0	-0.491	-0.120	-0.249	-0.190
SoutSup_6	0.455	0.554	0.503	0.000
style_3.0	0.398	-0.563	-0.288	0.000
smoking	0.324	0.255	0.543	0.000
origines_ethniques_4.0	0.267	0.371	0.635	0.000
maladies_15	-0.197	-0.166	-0.648	0.000
consult_who_5	0.186	0.198	0.174	0.000
origines_ethniques_2.0	-0.169	-0.179	-0.222	0.000
style_8.0	-0.150	-0.004	0.696	0.000
LatDec_3	-0.138	0.051	-0.552	0.000
issue_ai_data_3_4.0	-0.109	-0.249	-0.435	-0.576
chronotype_2.0	0.104	0.088	0.112	0.000
married_5.0	0.100	0.152	0.201	0.000
travail_domaine_8	0.045	0.096	0.142	0.000
maladies_19	0.042	0.131	0.073	0.000
travail_domaine_1	0.037	-0.179	-0.177	0.000
travail_domaine_3	0.033	0.000	0.076	0.000
car_model_4.0	0.031	-0.057	0.006	0.000
travail_domaine_10	-0.028	-0.037	-0.040	0.000
consult_who_3	-0.026	0.006	0.056	0.000
travail_domaine_2	0.025	0.245	0.242	0.000
maladies_22	0.013	0.008	0.102	0.000
travail_domaine_6	0.003	0.000	0.000	0.000
maladies_21	-0.001	0.000	0.170	0.000
maladies_20	0.000	0.000	-0.044	0.000
consult_who_6	0.000	0.000	-0.000	0.000
maladies_16	0.000	0.038	0.049	0.000

Python modeling

The complete Python code used in this thesis comes from a shared project repository and consists of several interconnected scripts and modules. These files depend on one another to perform data preprocessing, model training, hyperparameter tuning, performance evaluation, and explainability analyses. For this reason, presenting the entire code in the thesis would not be useful for the reader.

Instead, the following subsections present selected parts of the code that are particularly important for understanding the methodological implementation of this work. These excerpts illustrate how the machine learning models were defined and evaluated, as well as how the LIME and SHAP explanations were generated and analyzed. They therefore provide a clearer view of the main technical steps behind the results, without reproducing the full project structure.

Analyzed models

```python
# 1.2. Modelling
MODEL_LIST = [
    {
        "model_name": "mean_regressor",
        "param_grid": {},
    },
    {
        "model_name": "random_regressor",
        "param_grid": {},
    },
    {
        "model_name": "xgboost_regressor",
        "param_grid": {
            "imputer": [SimpleImputer(strategy="mean"), KNNImputer()],
            "scaler": [StandardScaler(), MinMaxScaler()],
            "regressor__max_depth": [3, 6, 9],
            "regressor__subsample": [0.5, 0.8],
        },
    },
    {
        "model_name": "ridge_regressor",
        "param_grid": {
            "imputer": [SimpleImputer(strategy="mean"), KNNImputer()],
            "scaler": [StandardScaler(), MinMaxScaler()],
            "regressor__alpha": [1.0, 2.0, 3.0, 4.0],
        },
    },
```

```python
    {
        "model_name": "lightgbm_regressor",
        "param_grid": {
            "imputer": [SimpleImputer(strategy="mean"), KNNImputer()],
            "scaler": [StandardScaler(), MinMaxScaler()],
            "regressor__n_estimators": [50, 100, 200],
            "regressor__learning_rate": [0.01, 0.1, 0.2],
            "regressor__max_depth": [-1, 3, 6],
            "regressor__num_leaves": [15, 31, 50],
        },
    },
    {
        "model_name": "catboost_regressor",
        "param_grid": {
            "imputer": [SimpleImputer(strategy="mean"), KNNImputer()],
            "scaler": [StandardScaler(), MinMaxScaler()],
            "regressor__depth": [4, 6, 8],
            "regressor__learning_rate": [0.01, 0.1, 0.2],
            "regressor__iterations": [50, 100, 200],
            "regressor__early_stopping_rounds": [10],
        },
    },
    {
        "model_name": "stacking_regressor",
        "param_grid": {
            "imputer": [SimpleImputer(strategy="mean"), KNNImputer()],
            "scaler": [StandardScaler(), MinMaxScaler()],
            "regressor__estimators": [
                [("catboost", CatBoostRegressor(depth=6, learning_rate=0.1, iterations=100)), ("ridge", Ridge(alpha=2.0))]
            ],
            "regressor__final_estimator": [
                CatBoostRegressor(depth=6, learning_rate=0.1, iterations=100),Ridge(alpha=2.0),
            ],
            "regressor__cv": [3],
        },
    ]
```

Global SHAP

```
# - SHAP feature importances

# Preprocess training data (everything before the regressor in the pipeline)
X_train_preprocessed = best_model[:-1].transform(X_train)
regressor = best_model.named_steps["regressor"]

# Number of samples to explain
n_samples = min(100, X_train_preprocessed.shape[0])
X_explain = X_train_preprocessed[:n_samples]

try:
    # Try the unified SHAP explainer (auto-selects best explainer)
    explainer = shap.Explainer(regressor, X_train_preprocessed, feature_names=selected_features)
    shap_values = explainer(X_explain).values
except Exception:
    # Fallback to KernelExplainer if auto explainer fails (slower)
    background_size = min(50, X_train_preprocessed.shape[0])
    background = shap.sample(X_train_preprocessed, background_size, random_state=42)
    explainer = shap.KernelExplainer(lambda x: regressor.predict(x), background)
    shap_values = np.array(explainer.shap_values(X_explain))
    # KernelExplainer.shap_values may return list for multilabel; handle that
    if isinstance(shap_values, list):
        shap_values = np.array(shap_values[0])

# Ensure shap_values is (n_samples, n_features)
if shap_values.ndim == 3:
    # e.g., (n_samples, n_outputs, n_features) -> take first output
    shap_values = shap_values[:, 0, :]

# Mean absolute SHAP value per feature
mean_abs_shap = np.mean(np.abs(shap_values), axis=0)
feature_coeff_dict = dict(zip(selected_features, mean_abs_shap))
sorted_feature_coeff_dict = dict(
    sorted(feature_coeff_dict.items(), key=lambda item: item[1], reverse=True)
)
```

Selection of the analyzed example

```
# Charger les données d'exemple pour éviter le biais du training set
df_example = pd.read_csv(
    ml_run_path / C.DEPLOY_FOLDER_NAME / C.EXAMPLE_ANSWERS_FILENAME,
    index_col=C.ATTRIBUTE_ID_COL
)

# Créer les features pour les exemples (exactement comme dans predict_for_example)
df_features_example = create_features_for_example(df_example, ml_run_path)

# Sélectionner les features utilisées dans le modèle
assert all(col in df_features_example.columns for col in selected_features)
X_example = df_features_example[selected_features].values

# Appliquer les mêmes transformations que le training set (imputer + scaler)
X_example_preprocessed = best_model[:-1].transform(X_example)

# Utiliser le premier exemple pour l'explication locale
X_local = X_example_preprocessed[0:1]
```

Local SHAP

```
# Calcul des valeurs SHAP localement pour l'échantillon choisi
shap_values_local = None
try:
    # Essayer TreeExplainer pour modèles d'arbres (ex. CatBoost, XGBoost, LightGBM)
    explainer = shap.TreeExplainer(regressor)
    shap_out = explainer.shap_values(X_local)
    shap_values_local = np.array(shap_out)
except Exception:
    try:
        # Essayer l'explainer générique (auto)
        explainer = shap.Explainer(regressor, X_train_preprocessed, feature_names=selected_features)
        shap_out = explainer(X_local)
        shap_values_local = np.array(shap_out.values)
    except Exception:
        # Dernier recours : KernelExplainer (lent mais robuste)
        background_size = min(50, max(1, X_train_preprocessed.shape[0]))
        background = shap.sample(X_train_preprocessed, background_size, random_state=42)
        explainer = shap.KernelExplainer(lambda x: regressor.predict(x), background)
        shap_vals = explainer.shap_values(X_local)
        shap_values_local = np.array(shap_vals)
        if isinstance(shap_values_local, list):
            shap_values_local = np.array(shap_values_local[0])

# Normaliser la forme en vecteur 1D de longueur n_features
if shap_values_local is None:
    # Si tout a échoué, définir contributions à zéro
    shap_values_local = np.zeros(len(selected_features))
elif shap_values_local.ndim == 3:
    # (1, n_outputs, n_features) -> prendre premier output
    shap_values_local = shap_values_local[0, 0, :]
elif shap_values_local.ndim == 2:
    # (1, n_features) -> aplatir
    shap_values_local = shap_values_local[0, :]

# Dictionnaire des contributions signées triées par contribution absolue décroissante
feature_local_contrib = dict(zip(selected_features, shap_values_local))
sorted_feature_coeff_dict = dict(
    sorted(feature_local_contrib.items(), key=lambda item: abs(item[1]), reverse=True)
)
```

LIME

```
    # Calcul d'une explication LIME locale pour l'échantillon choisi
# On utilise les données d'entraînement prétraitées comme base de référence
lime_explainer = LimeTabularExplainer(
    training_data=X_train_preprocessed,
    feature_names=selected_features,
    mode="regression",
    discretize_continuous=True,
    random_state=42,
)

# LIME explique une instance 1D, donc on prend X_local[0]
lime_exp = lime_explainer.explain_instance(
    data_row=X_local[0],
    predict_fn=regressor.predict,
    num_features=len(selected_features),
)

# Récupérer les contributions LIME sous forme de dictionnaire
# exp.as_map()[1] correspond généralement à la sortie en régression
lime_weights = lime_exp.as_map()[1]

feature_local_contrib = {
    selected_features[feature_idx]: weight
    for feature_idx, weight in lime_weights
}

# Ajouter les features absentes de l'explication avec une contribution nulle
# pour garder le même format que précédemment
for feature in selected_features:
    feature_local_contrib.setdefault(feature, 0.0)

# Dictionnaire des contributions signées triées par contribution absolue décroissante
sorted_feature_coeff_dict = dict(
    sorted(
        feature_local_contrib.items(),
        key=lambda item: abs(item[1]),
        reverse=True,
    )
)
```

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm