IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Minimum Description Length and Multi-Criteria Decision Analysis in Predictive Modelling

Petr Silhavy[1], Kateřina Hlaváčková-Schindler[2], Radek Silhavy[1]
[1]Tomas Bata University in Zlin, Faculty of Applied Informatics, Zlin, Czech Republic
[2]Data Mining and Machine Learning Research Group, Faculty of Computer Science, University of Vienna, Vienna, Austria

Corresponding authors: Radek Silhavy (radek@silhavy.cz), K. Hlaváčková-Schindler (katerina.schindlerova@univie.ac.at)

Tomas Bata University in Zlín, Faculty of Applied Informatics, Project number: RO30246061025/2102

ABSTRACT Accurate model selection is essential in predictive modelling across various domains, significantly impacting decision-making and resource allocation. Despite extensive research, the model selection process remains challenging. This work aims to integrate the Minimum Description Length principle with the Multi-Criteria Decision Analysis to enhance the selection of forecasting machine learning models. The proposed MDL-MCDA framework combines the MDL principle, which balances model complexity and data fit, with the MCDA, which incorporates multiple evaluation criteria to address conflicting error measurements. Four datasets from diverse domains, including software engineering (effort estimation), healthcare (glucose level prediction), finance (GDP prediction), and stock market prediction, were used to validate the framework. Various regression models and feed-forward neural networks were evaluated using criteria such as MAE, MAPE, RMSE, and Adjusted R^2. We employed the Analytic Hierarchy Process (AHP) to determine the relative importance of these criteria. We conclude that the integration of MDL and MCDA significantly improved model selection across all datasets. The cubic polynomial regression model and the multi-layer perceptron models outperformed other models in terms of AHP score and MDL criterion. Specifically, the MDL-MCDA approach provided a more nuanced evaluation, ensuring the selected models effectively balanced complexity and predictive accuracy.

INDEX TERMS Multicriteria Decision Analysis; Minimum Model Length; Machine Learning; Model Selection Prediction; MDL-MCDA

I. INTRODUCTION

Accurate model selection is essential in predictive modelling across various domains. The efficacy of predictive models influences decision-making processes and resource allocation. Despite extensive studies comparing multiple predictive models, the model selection approach still needs to be explored. Model selection is intrinsically tied to the objectives of prediction and understanding, with its essence captured through the formalisation of loss and risk, as declared by Petropoulos et al. [1] and by Friedman [2]. The issues of model selection lie in navigating through the complex relationship between independent variables and the dependent variable underpinned by both observable and unobservable factors. The literature identifies two broad categories of variables influencing dependent variables: explanatory variables, which are observable, and unobservable variables, which include factors such as measurement errors or omitted independent variables. This issue has been discussed in various problem domains.

A. HOW MODELS CAN BE SELECTED: MAIN ISSUES

Model selection is a crucial challenge within all prediction tasks, bridging the gap between theoretical constructs and practical applications. This section describes the core aspects of model selection, covering the fundamental issues, model evaluation and construction methodologies, and the selection process. The multicriteria approach, assumptions underlying model selection strategies, and the interplay between explanatory variables and unobservable factors influencing the outcome variable are essential to our discussion. The evaluated datasets are often more complex because they contain more features, allowing more independent variables to be used. Issues in model selection are mainly related to selecting a model that will fit data well, keep a complexity level low, and provide a reasonable, accurate prediction. For each prediction task, the following is to be evaluated [3]:

- Complexity level – The existing methods of assessing model quality are often based on assumptions of randomness of variables and may, therefore, be sensitive

to extreme values. On the contrary, some methods make few assumptions about randomness, but their inherent generality may alter their results. This means that while assumption-based methods can be very accurate under ideal conditions, they may fail with non-ideal data, whereas assumption-light methods are more flexible but can sometimes offer less precise insights.

- Class of models for a specific system – It is also important to use a good set of predictive models and select the most relevant method for model construction.
- Evaluation criteria – Selecting the most relevant evaluation criteria for prediction models is crucial. A model evaluation criterion, based on its distance to the theoretical quantity, assesses the performance in predicting a model. Also, criteria can often be in conflict.

The prediction task is challenging due to unknown relationships between the variables involved. A common approach is to create multiple models that represent these relationships differently. The task then becomes evaluating and comparing these models to select the best one, where the "best" is task-specific. Four benchmark problem domains were chosen to validate the proposed evaluation methods: effort estimation, predicting glucose levels, gross domestic product (GDP) prediction, and stock price prediction. These domains require accurate predictions in both technical and financial fields and the versatility and robustness of the proposed MCDA-MDL evaluation framework is demonstrated in our work. The first prediction task is in software engineering, focusing on predicting software development efforts. This task is crucial for project management because accurate estimates help to plan, budget, and allocate resources. Simple and accurate predictive models are valuable as they are easier for project managers and stakeholders to understand and use, ensuring better project control and success. The second prediction task is from the medical and health science, explicitly predicting glucose levels. Accurate glucose level predictions are essential for managing diabetes, as they help to monitor and maintain optimal glucose levels and prevent complications. This task represents a broader challenge in medical research, where accurate predictions are necessary for effective patient care and treatment planning.

The third prediction experiment is related to the gross domestic product prediction. Those predictions are essential for company financial planning and life cost prediction. Knowing the gross domestic product prediction is mandatory for many businesses and public administration.

The fourth domain is related to stock market prediction. This was included as being a typical representation of the time series. Also, this is an important task in economic and financial analysis.

By choosing these four significant problem domains, the study aims to show the versatility and robustness of the proposed MCDA-MDL evaluation framework over data science applications. The software engineering, medical and financial analysis/economic tasks highlight the need for practical and easy-to-use models in each field.

B. OBJECTIVES OF THE WORK

Model selection simplifies the process by reducing the number of possible models to a limited set. However, it remains a challenging problem because it requires defining what makes a good model and how to measure its quality. These definitions should align with the primary goal of the study. Although this seems straightforward, in practice, the methods used to create and evaluate models often need to align better with the study's objectives.

To address the challenges of model selection in the presence of conflicting error measurements, we propose the integration of Minimum Description Length (MDL) and Multi-Criteria Decision Analysis (MCDA).

The MDL principle [4], [5] helps balance the model's complexity with its ability to fit the data. By minimizing the minimal description length, MDL provides a robust way to prevent overfitting and select models that generalize well to new data. Integrating the MCDA approach is essential when error measurements conflict. MCDA helps to incorporate an error score, which fuses more than one error criterion. MCDA score can be understood as a goodness-of-fit part of MDL. By combining MDL and MCDA, we can enhance the model selection process, ensuring that the selected model fits the data well and effectively meets the task's objectives. This integrated approach provides a structured framework to navigate the complexities of model evaluation and selection.

C. RESEARCH QUESTIONS

For this work, the following research questions have been set:

- **RQ1:** How does the Minimum Description Length (MDL) and Multi-Criteria Decision Analysis (MCDA) integration affect predictive model selection?
- **RQ2:** What advantages does the MDL-MCDA have compared to the MDL-RSS[1] in predictive model selection?

D. MAIN CONTRIBUTIONS OF THE WORK

In this paper, we address critical challenges in predictive modeling and model selection by introducing a novel methodological framework that integrates the strengths of the Minimum Description Length (MDL) principle with Multi-Criteria Decision Analysis (MCDA). While traditional MDL relies on Residual Sum of Squares (RSS) as a measure of goodness-of-fit, this approach often falls short in scenarios where multiple, conflicting error criteria must be balanced, particularly in complex, real-world datasets. By integrating MDL with MCDA, we extend the scope of model evaluation beyond a single error metric, allowing for a more robust and nuanced assessment that accounts for multiple evaluation criteria. This innovation not only enhances the reliability of model selection but also addresses critical gaps in traditional MDL approaches. Below, we detail the specific contributions that

[1] MDL-RSS is the common MDL having the residual sum of squares at the goodness-of-fit criterion.

underscore the novelty and practical value of our proposed framework:

- **MDL and MCDA Integration:** Introduces a novel integration of Minimum Description Length (MDL) with Multi-Criteria Decision Analysis (MCDA) to improve predictive model selection by resolving conflicting error measurements.
- **Comparative Analysis:** Evaluates MDL-AHP vs. MDL-RSS integrations for better handling of complex datasets, showing practical benefits.
- **Domain Applications:** Assesses the methodology across multiple domains.
- **Benchmark Datasets:** Validates the framework using datasets from software engineering, medical, and finance domains.
- **Enhanced Selection Framework:** Demonstrates that MDL-MCDA improves model selection by balancing complexity and accuracy.
- **Impact of MDL-MCDA:** MDL-MCDA outperforms traditional MDL or MCDA in model selection across datasets.
- **MDL-MCDA vs. MDL-RSS:** Shows MDL-MCDA selects models with better generalization compared to MDL-RSS.

These contributions advance the understanding and implementation of model selection methodologies, offering to apply MDL with MCDA to various domains requiring precise and reliable predictive modelling.

E. PAPER ORGANISATION

The rest of the paper is organised as follows. Section 2 provides a comprehensive overview of existing research and methodologies related to model selection, MDL, MCDA, *AHP*, and RSS. Section 3 details the methods used in our work, including data preparation, model implementation, evaluation measures, and the integration of MDL and MCDA. Section 4 presents the results of the experiments using various datasets and predictive models. It includes a comparison of the performance of different models based on *AHP*, MDL with *AHP*, and MDL with RSS. Section 5 discusses the implications of the results, the effectiveness of the integrated approach, and its applicability to different problem domains. Finally, Section 6 summarizes the study's main findings, highlights the contributions, and suggests directions for future research.

II. RELATED WORK

A wide range of viable prediction models are available across different industries, making it difficult to determine the optimal one, especially when faced with conflicting error measures. The Minimal Description Length (MDL) was introduced to address this issue in [4]. MDL is an alternative to the Akaike Information Criterion (AIC), which was introduced as a recognised method for automatic model selection [6]. While the Akaike Information Criterion (AIC) is highly efficient in selecting models within the same class and comparing non-nested models, such as linear and non-linear models, it cannot automatically choose models from different prediction model classes, such as exponential smoothing and autoregressive models. To address this limitation, the Bayesian information criterion (BIC) from Schwarz was introduced, which, in the same vein as AIC, evaluates the fit of the data with a complexity penalty. However, the BIC imposes a more substantial penalty for complexity than the AIC. Nevertheless, this method still requires further development to assess models within the same class. Villegas et al. [7] suggest employing support vector machines (SVM) to identify the most suitable prediction model from a range of alternatives, given that model variables (such as the degree of accuracy and the fitted parameters) may change over time. The researchers discovered that utilising SVM leads to a greater overall predictive accuracy. Ghobbar and Friend [8] devised a predictive error forecasting technique for assessing demand prediction models in the airline manufacturing sector based on their factor levels. They employed mean absolute percentage error (MAPE) as the criterion for evaluation but did not account for hybrid prediction models that incorporate personal information. Oh and Morzuch [9] assessed eight demand prediction models using six performance measures that evaluate bias and forecast error, including MAPE, MAE, RMSE, AIC, and BIC. Their study revealed that the choice of prediction model varied based on the performance measures employed. Taylor and McSharry [10] evaluated six distinct prediction models to estimate electricity demand across ten European countries. They used MAPE and MAE as evaluation measures and discovered that the rankings generated conflicting outcomes, except for the top-performing model, which consistently ranked first. Petropoulos et al. [1] and Han et al. [11] investigated the use of subjective expert judgment in prediction model selection, revealing that the chosen models outperformed those selected through AIC based on evaluation measures such as MAE, MAPE and MASE. Furthermore, it has been shown that collective judgment is superior to a single decision and statistical selection methods. Davydenko and Fildes [6], for instance, explored the effectiveness of MAPE and median average percentage error (MdAPE) in assessing judgmental adjustments to statistical prediction. They concluded that relying solely on MAPE to determine a model's performance is insufficient due to inconsistent results between MAPE and other error measures. The study suggests that future research should develop an approach for selecting the optimal model when evaluating multiple error measures, particularly in the face of conflicting results. Multiple-criteria decision analysis (MCDA) is a widely-used approach for addressing complex problems involving multiple, often conflicting, objectives [12]. Selecting predictive models using MCDA can be particularly useful when different error measures, such as mean squared error and mean absolute error, provide conflicting guidance on the optimal model. Comparing *AHP* and TOPSIS, two prominent MCDA methods, we consider how they can be applied in this model selection context. The Analytic Hierarchy Process (AHP) [13] is a

structured technique for organizing and analyzing complex decisions. *AHP* involves decomposing a problem into a hierarchy of goals, objectives, and alternatives and then using pairwise comparisons to derive priorities for the alternatives. In the model selection domain, *AHP* could be used to establish a hierarchy with the overall goal of minimizing prediction error, with sub-objectives of minimizing MSE, MAE, and potentially other relevant measures. Each candidate model would then be evaluated against these criteria, with *AHP* providing a composite score to guide the final model selection [14] [15] [16] [17]. The perceptron, a fundamental building block of neural networks, has also been explored for effort estimation. A study by [18] demonstrated the potential of perceptron-based models to capture non-linear relationships, characteristic for effort estimation problems and the potential to improve traditional estimation techniques. While neural network and deep learning models have shown promising results, their performance is heavily dependent on the quality and characteristics of the input data. Proper feature engineering, data preprocessing, and hyperparameter tuning are crucial for achieving reliable and accurate effort estimation using these advanced techniques. Hyperparameter optimization can significantly impact the model's predictive capabilities and generalisation, such as the number of hidden layers, neurons, and the learning rate.

Neural networks and deep learning models promise to improve software effort estimation. Their ability to model complex, non-linear relationships in data makes them well-suited for this task. Continued advancements in neural network architectures, training algorithms, and hybrid modelling approaches will likely enhance their accuracy and applicability in software engineering.

Various methods for evaluating prediction models in different domains, primarily using error measures and information criteria like AIC and BIC. However, employing AIC and BIC to assess models restricts the comparison to models within the same class. Moreover, further research is needed to determine an appropriate approach to evaluating multiclass demand prediction models based on several interdependent error measures and to select the best model based on the simultaneous use of multiple error measures [19].

III. METHODOLOGY

A. RESEARCH DESIGN

This work evaluates the integration of minimum description length (MDL) and multi-criteria decision analysis (e.g. *AHP*) in selecting predictive models. To achieve this, a series of steps were taken during experimental work. This involves data preparation, model implementation, evaluation, and comparison.

We employ four datasets covering software engineering, medical and financial problem domains. These cover several domains and sizes and are also a combination of natural and synthetic samples. The considered model classes include multiple linear regression (MLR), Ridge regression, Lasso regression, Elastic net regression, quadratic and cubic regression (i.e. polynomial regression with degrees 2 and 3), and a feed-forward neural network (FF-NN) with various configurations, which will be specified.

The performance of these models will be evaluated using criteria: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Adjusted R-squared (adjR^2), Prediction at 25% (Pred(0.25)), and Weighted Quantile Loss (WQL). Additionally, we will apply the Analytic Hierarchy Process (AHP) and Minimum Description Length (MDL) principles to aid in model selection. The procedure will involve several steps (Figure 1).

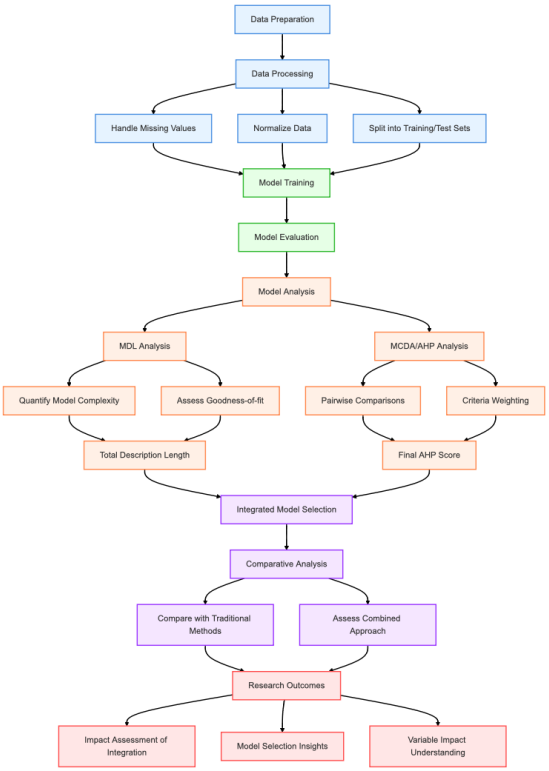

FIGURE 1: Model Design Flowchart.

First, a dataset is prepared and processed by handling missing values, normalising data, and splitting them into training and test sets. Then, the predictive models are trained using the training data. After training, they evaluate each model using the specified criteria on testing data. The MDL principle will help quantify the complexity and goodness-of-fit for each model, focusing on the total description length, which includes the model structure, parameters, and data encoding. MCDA, specifically *AHP*, will evaluate models based on multiple criteria. This involves making pairwise comparisons to determine the relative importance of each criterion in the final score for model selection.

Finally, the models selected using traditional methods will be compared with those chosen through the integrated MDL and MCDA approach. This comparison will determine if the combined approach improves predictive accuracy and model simplicity.

The expected outcomes of this work include identifying the impact of integrating MCDA methods (e.g. *AHP*) for MDL and comparing to selection using *AHP* only, or MCDA with RSS. Moreover, we obtain insights into how MDL and MCDA can improve model selection and understand the impact of different variables on model performance.

B. EVALUATING MEASURES

The evaluation and comparison of models involve a detailed analysis of various models' performance, focusing on their ability to predict or explain the dependent variable accurately. Let us consider a sample $D = \{(x_i, y_i), i = 1, \ldots, n\}$, of variable values y_i and $\hat{y}_i$, where y_i represents the actual value and $\hat{y}_i$ is the predicted value.

Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy of a forecasting method, expressing the accuracy as a percentage. It is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100. \tag{1}$$

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are common evaluation measures assessing the average magnitude of prediction errors. MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

and RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3}$$

The Median Absolute Percentage Error (MdMAPE) provides a robust measure by focusing on the median of the percentage errors, defined as:

$$\text{MdMAPE} = \text{median}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\right) \tag{4}$$

The Adjusted Coefficient of Determination (Adjusted $\mathbf{R^2}$) is an enhancement of the regular $\mathbf{R^2}$ metric that adjusts for the number of predictors in the model. It provides a more accurate measure of goodness of fit than R^2 by considering model complexity. Adjusted $\mathbf{R^2}$ is defined as:

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1} \tag{5}$$

where n is the number of observations, k is the number of predictors, and R^2 is the coefficient of determination on set D.

Pred(0.25) evaluates the proportion of predictions that fall below a specified error threshold, such as 25%. It is useful for assessing the overall model's predictive accuracy within an acceptable error range. Pred(0.25) is calculated as

$$\text{Pred}(0.25) = \frac{1}{n}\sum_{i=1}^{n} I\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| < 0.25\right) \tag{6}$$

where I is an indicator function that equals 1 if the condition is true and 0 otherwise.

The Weighted Quantile Loss (WQL) [20] measures how well a predictive model performs across different quantiles of the target variable's distribution. It is particularly useful in scenarios where it is important to understand the model's performance across various data distribution segments. The WQL is given by:

$$\text{wQL}(\tau) = \frac{\sum_{i=1}^{N} L_\tau(y_i, \hat{y}_i(\tau))}{\sum_{i=1}^{N}|y_i|} \tag{7}$$

where τ is the quantile level (e.g., quartils), y_i is the observed value at the i-th data point, and $\hat{y}_i(\tau)$ is the predicted quantile value at the i-th data point for the quantile level τ. The quantile loss function, $L_\tau(y_i, \hat{y}_i(\tau))$, is defined as $L_\tau(y_i, \hat{y}_i(\tau)) = (\tau - \mathbf{1}\{y_i < \hat{y}_i(\tau)\})(y_i - \hat{y}_i(\tau))$, where $\mathbf{1}\{y_i < \hat{y}_i(\tau)\}$ is an indicator function that equals 1 if $y_i < \hat{y}_i(\tau)$ and 0 otherwise.

1) Discussion on Evaluation Measures

Model performance is assessed using various criteria: Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Median Absolute Percentage Error (MdMAPE), Adjusted R^2, and Predictions Level (Pred(0.25)). Each criterion offers a distinct perspective on model evaluation, capturing different facets of accuracy, robustness, or complexity. However, these measures may occasionally produce conflicting results, necessitating careful interpretation.

The selection of these criteria reflects their ability to balance accuracy, robustness to outliers, and the trade-off between model fit and complexity.

Accuracy-focused measures:

- MAPE measures percentage errors, providing an intuitive view of relative accuracy for stakeholders.
- MAE averages error magnitudes, offering a straightforward overall accuracy measure without outlier bias.
- RMSE highlights large errors, useful for significant deviations but sensitive to outliers, potentially conflicting with MAPE.

Robustness against outliers:

- MdMAPE captures median percentage errors, ensuring robustness to outliers and complementing RMSE and MAE.

Model fit and complexity:

- Adjusted R^2 measures variance explained, accounting for predictors to balance fit and complexity.
- Pred(0.25) measures predictions within 25% error, prioritizing consistent accuracy over complexity metrics.

Balancing these criteria is crucial for developing robust models. Measures like MAPE, MAE, RMSE, MdMAPE, and Pred(0.25) focus on predictive accuracy, while complexity-oriented measures like Adjusted R^2 provide insights into generalizability. By evaluating multiple metrics, a comprehensive understanding of the model's strengths and weaknesses emerges.

Evaluation measures can be grouped based on whether they should be minimized, maximized, or zeroed for optimal performance:

Minimization Criteria:

- Mean Absolute Percentage Error (MAPE): emphasizes relative prediction errors.
- Mean Absolute Error (MAE): captures average error magnitudes.
- Root Mean Squared Error (RMSE): penalizes larger errors, highlighting extreme deviations.
- Median Absolute Percentage Error (MdMAPE): offers robustness to outliers.

Maximization Criteria:

- Adjusted Coefficient of Determination (R^2): balances variance explanation and model complexity.
- Proportion of Predictions Below 25% Error (Pred(0.25)): emphasizes practical predictive accuracy.
- Weighted Quantile Loss: ensures balanced performance across quantiles.

This diverse set of evaluation criteria ensures the model is both accurate and generalizable, meeting practical needs while avoiding overfitting or overemphasis on specific error types.

C. MINIMUM DESCRIPTION LENGTH

The Minimum Description Length (MDL) principle is a formal method of inductive inference that balances the model complexity and goodness of fit. This principle is rooted in information theory and aims to avoid overfitting by penalising model complexity. MDL was introduced by Rissanen [4] and further developed by Rissanen, Barron, Yu in e.g. [21]–[23], and by Grünwald and Roos in [5], [24].

MDL is based on the idea that the best model for a given set of data is the one that allows for the shortest overall description of the data and the model itself. The total description length is the sum of the data and model encoding lengths. Mathematically, the total description length $L(D, M)$ can be expressed as:

$$L(D, M) = L(M) + L(D|M) \tag{8}$$

where $L(M)$ is the length of the description of the model M and $L(D|M)$ is the length of the description of the data D given the model M.

MDL prefers models that balance simplicity (short model description) and accuracy (short data description given the model) in the sense that the best model D is minimizing Eq. (8). This approach penalizes more complex models unless they significantly improve the data fit. MDL can bring advantages where overfitting is a concern and model interpretability and simplicity are valued. MDL helps select models that generalise well to new data by penalizing model complexity.

MDL focuses on the total length of encoding both the model and the data, ensuring that the model chosen is the one that best compresses the data. This means MDL inherently balances model fit and complexity by minimizing the information required to describe the model and the data it explains. Unlike AIC and BIC, which are derived from statistical considerations, MDL directly addresses the issue of overfitting by penalizing unnecessarily complex models, thus often leading to models that generalize better to new data. This makes MDL a robust criterion for selecting models that are not only accurate but also parsimonious, enhancing predictive performance and interpretability.

1) Two-Part MDL Codes

The minimal description length as defined by Eq. (8) is in the literature called a two-part MDL. We point in the beginning, since MDL is a principle, there can be various encodings of the models from a class of models and thus there can be various MDL functions corresponding to the general scheme from Eq. (8).

We will now come to a more formal explanation of the MDL principle and its encodings. First we briefly explain the orginal theory from Rissanen [4], [22] and his followers [5], [24] as it was developed for the case that a conditional probability distribution $p(y|x)$ is known. Secondly we explain MDL when we only know about the model, from which the data are generated, that it is a member of a class of functional models [25].

2) Encoding of Models with Known Probability of the Data Deneration Process

We define a model for the prediction problem as a conditional probability distribution $p(y|x)$ over and input space X, i. e. in other words, $\sum_{y\in Y} p(y|x) = 1$ (where the output space Y can be theoretically also an infinite). A model class is a set of models depending on a parameter vector $\boldsymbol{\theta}$, i.e. $M = \{p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Usually, $\boldsymbol{\Theta}$ is a subset of a multivariate Euclidean space. Shannon in [26] proved the following fundamental statement in information theory, known under the name Shannon-Huffman code. If a sender and receiver agreed in advance on a model p and both know the input $x_i, i = 1, \ldots, n$ then there exists code to transmit the values $y_i, i = 1, \ldots, n$ losslessly with codelength (up to at most one bit on the whole sequence)

$$L_p(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{n} \log_2 p(y_i|x_i) \tag{9}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$ is a shortened notation for set $y_1, \ldots, y_n$, $x_1, \ldots, x_n$, respectively (which are from Y, X respectively). The one additional bit in the Shannon-Huffman code is present only once for the whole data set [27] and with large data sets is negligible. Thus it will be omitted from the encodings.

We do no need to know the practical implementation of compression algorithms but we consider only the theoretical bit length of their associated encodings. We want to measure the amount of information contained in the data, and how it is represented by the model. So we will directly work with codelength functions. Probability distribution function p can

be understood as the data generating process and in general it is not known but can be approximated from the data.

To quantify the complexity of the computational models for prediction (and in general for a supervised learning problem) can be done e.g. by parameter counting. An information-theoretic way to use the Occam razor principle in terms of the simplest model with a good generalization is the minimum description length (MDL), introduced by Rissanen [4] and further developed by Rissanen, Barron, Yu in e.g. [21]–[23], and by Grünwald and Roos in [5], [24]. Encodings in which the parameters of a model are at first transmitted to the receiver and then the data using these parameters are encoded, have been called two-part codes and introduced by Grünwald [5].

Let $L_{param}(\boldsymbol{\theta})$ be any encoding scheme for parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and let $\boldsymbol{\theta}^*$ be any parameter. The corresponding two-part codelength is

$$\begin{aligned} L_{\boldsymbol{\theta}^*}(\boldsymbol{y}|\boldsymbol{x}) &= L_{param}(\boldsymbol{\theta}^*) + L_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{y}|\boldsymbol{x}) \\ &= L_{param}(\boldsymbol{\theta}^*) - \sum_{i=1}^{n} \log_2 p_{\boldsymbol{\theta}^*}(y_i|x_i). \end{aligned} \tag{10}$$

$\sum_{i=1}^{n} \log_2 p_{\boldsymbol{\theta}^*}(y_i|x_i)$ is called the goodness-of-fit. The objective is to find $\boldsymbol{\theta}^*$ at the minimum of (10) over all parameterizations.

3) Encoding of Models with Known Functional Class of the Data Generating Process

When p is known, it is clear that the minimum of (10) is equivalent to the maximum likelihood estimate (MLE). However, what makes the MDL principle so generic is that it can be generalized to the functional cases, i.e. instead of p probability, f as a general function can be considered about which is only known to be a member of a class of candidate models, see e.g. [25]. It means that about the model, from which the data are generated, is only known to be a member $f_l(.|\boldsymbol{\theta}_l)$ of a class of models

$$\begin{aligned} M = \{f_l(.|\boldsymbol{\theta}_l), \boldsymbol{\theta}_l \in \boldsymbol{\Theta}_l, \theta_{lj} \sim \pi_{lj}(\theta_{lj}), \\ l = 1, \ldots, m, j = 1, \ldots, k_l\} \end{aligned} \tag{11}$$

where m is the number o models in M, $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lk_l})$ is a k_l-dimensional parameter vector associated with f_l and $\boldsymbol{\Theta}_l$ is a parameter space for $\boldsymbol{\theta}_l$. $\pi_{lj}(\theta_{lj})$ is introduced merely to simplify the encoding process as an artificial device to minimize the description length. It is assumed that every f_l is known except for $\boldsymbol{\theta}_l$, and that different f_l may have different number of parameters k_l. Given a set of observed data, the goal is to find the "true" f_l from M as well as to estimate the parameter $\boldsymbol{\theta}_i$ associated with it. In this sense is (10) replaced by

$$L(\boldsymbol{y}) = L(\hat{\boldsymbol{\theta}}_l) + L(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_l) \tag{12}$$

where $L(\hat{\boldsymbol{\theta}}_l)$, $L(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_l)$ are code lengths for encoding $f_l(.|\hat{\boldsymbol{\theta}}_l)$ and "$\boldsymbol{y}$ conditioned on $f_l(.|\hat{\boldsymbol{\theta}}_l)$" respectively. $L(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_l)$ is called the goodness-of-fit.

Rissanen in [21] proved that if $\hat{\theta}_{lj}$ is an MLE computed from n_j data points and if n is large, then the precision of θ_{lj} can be effectively encoded with $\frac{1}{2}\log_2 n_j$ bits. Rissanen derived a well-known form when all the parameters θ_{lj} are to be estimated by using all data points of size n. For subset selection in regression analysis, based on [22] it is

$$MDL(k_l) = -\log_2 f_l(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_l) - \sum_{j=1}^{k_l} \log_2 \pi_{lj}(\hat{\theta}_{lj}) + \frac{k_l}{2}\log_2 n. \tag{13}$$

where k_l is the number of the regressors. Moreover, for n large, the choice of $\pi_{lj}(\hat{\theta}_{lj})$ is relatively unimportant, as the resting summands in (13) are dominating [22]. So in practice for high n, term $\pi_{lj}(\hat{\theta}_{lj})$ can be omitted for MDL.

4) MDL with Multi-Objective Goodness-of-Fit

We propose to replace the goodness-of-fit measure, which are commonly used in the MDL literature, namely *MAE*, *MAPE* *RMSE* etc. by the multiobjective goodness-of-fit measure. We will utilize this idea of both probabilistic and functional representation of models described above.

If the probability function p is known: The second part in Eq. (10) is a goodness-of-fit of the model $p_{p_{\boldsymbol{\theta}^*}}$ on data set D. In this paper, we replace in the value $L_{p_{\boldsymbol{\theta}^*}}(\boldsymbol{y}|\boldsymbol{x}) := \sum_{i=1}^{n} -\log_2 p_{\boldsymbol{\theta}^*}(y_i|x_i)$ by

$$L_{\Phi_{\boldsymbol{\theta}^*}}(\boldsymbol{y}|\boldsymbol{x}) = -\log_2 \Phi_{\boldsymbol{\theta}^*}(\boldsymbol{y}|\boldsymbol{x}) \tag{14}$$

where $\Phi_{\boldsymbol{\theta}^*}(\boldsymbol{y}|\boldsymbol{x})$ is a multi-objective criterion, and similarly, as above, the first part $L_{param}(\theta^*)$ is the encoding of the selected model. The objective is to find θ^* at the minimum of (14) over all parameterizations.

If only function f_l is known, the goodness-of-fit in (13)

$$L(\boldsymbol{y}|\boldsymbol{x}) := -\log_2 f_l(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_l) = \sum_{i=1}^{n} -\log_2 f_l(y_i|\hat{\boldsymbol{\theta}}_l) \tag{15}$$

will be replaced analogically by (14).

5) Considered Machine Learning Methods

Our work incorporated three types of machine learning methods which we call model classes: multiple linear regression (LR), multiple linear regression with a penalization term (penLR), polynomial regression (polREG) up to degree 3, and feed-forward neural networks (FF-NN). In this work, multilayer perceptrons with two and three hidden layers are used.

In the following subsections, we construct the MDL descriptions of the above models for the multi-objective goodness-of-fit.

6) MDL for Linear Regression with Multi-Objective Goodness-of-Fit

Giurcăneanu et al. in [28] constructed several information-theoretic criteria for the variable selection by multiple linear regression assuming that the noise follows a Gaussian distribution. We will use their MDL derived from the stochastic

complexity [29]. However, we replace their goodness-of-fit with the multi-objective goodness-of-fit. We denote $k = |\boldsymbol{\gamma}|$ the number of non-zero values in the binary vector $\boldsymbol{\gamma}$, i. e. the number of regressors, and we can assume that $k > 0$. Let $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \in \mathbf{R}^{k\times 1}$ be the vector of the unknown regression coefficients within the $\boldsymbol{\gamma}$-subset. The matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is given by the columns of $\boldsymbol{X}$ that correspond to the $\boldsymbol{\gamma}$-subset and the regression equation is

$$\boldsymbol{y} = \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \varepsilon_{\boldsymbol{\gamma}}, \tag{16}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the dependent variable and $\varepsilon_{\boldsymbol{\gamma}}$ are Gaussian distributed with zero-mean and unknown variance $\tau_{\boldsymbol{\gamma}}$. Under the assumption that matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ has full-rank, the maximum likelihood (ML) estimates are

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\boldsymbol{X}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{y} \tag{17}$$

and

$$\hat{\tau}_{\boldsymbol{\gamma}} = \|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|_2^2/n \tag{18}$$

where $\hat{\tau}_{\boldsymbol{\gamma}}$ is a goodness-of-fit in Eq. (12) on regressor from $\boldsymbol{\gamma}$. Paper [28] evaluated MDL of these regressions with independent variables indexed by $\boldsymbol{\gamma}$ as functions depending on vector $\boldsymbol{y}$ and $\boldsymbol{\gamma}$ as

$$\begin{aligned} MDL^{LR}(\boldsymbol{y},\boldsymbol{\gamma}) = \frac{n-k}{2}\log_2\hat{\tau}_{\boldsymbol{\gamma}} + \frac{k}{2}\log_2\frac{\|\boldsymbol{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|_2^2}{n} \\ - \log_2\Gamma(\frac{n-k}{2}) - \log_2\Gamma(\frac{k}{2}) + \frac{n}{2}\log_2(n\pi) \end{aligned} \tag{19}$$

where Γ denotes the Euler integral of the second kind. In our MDL, we propose to replace $\hat{\tau}_{\boldsymbol{\gamma}}$ in (19) by a multi-objective criterion $\Phi_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}}(y_i|x_i)$, i.e.

$$\begin{aligned} MDL^{LR}_{\Phi}(\boldsymbol{y},\boldsymbol{\gamma}) = \frac{n-k}{2}\log_2(\frac{\hat{\Phi}_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}}(\boldsymbol{y}|\boldsymbol{x})}{n}) + \\ \frac{k}{2}\log_2\frac{\|\boldsymbol{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|_2^2}{n} \\ - \log_2\Gamma(\frac{n-k}{2}) - \log_2\Gamma(\frac{k}{2}) + \frac{n}{2}\log_2(n\pi). \end{aligned} \tag{20}$$

The objective is to find $\boldsymbol{\beta}^*$ at the minimum of (20) over all parameterizations $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and combinations of $\boldsymbol{\gamma}$.

7) MDL for penalized linear regression with multi-objective goodness-of-fit

We express the encoding of the penalization part in regression as $\frac{1}{2}\log_2\lambda$ for fixed values of MLE of $\hat{\beta}$. We use the same encoding of the regularization parameter for Lasso, Ridge and Elastic penalization. However, we stress that MDL minimization can be used only within the regression class with the same penalization type and not within all penalty types. Then

$$\begin{aligned} MDL^{penLR}(\boldsymbol{y},\gamma) = \frac{n-k}{2}\log_2\hat{\tau}_{\gamma} + \frac{1}{2}\log_2\lambda + \\ \frac{k}{2}\log_2\frac{\|\boldsymbol{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\gamma}\|_2^2}{n} - \log_2\Gamma(\frac{n-k}{2}) - \log_2\Gamma(\frac{k}{2}) + \\ + \frac{n}{2}\log_2(n\pi). \end{aligned} \tag{21}$$

and

$$\begin{aligned} MDL^{penLR}_{\Phi}(\boldsymbol{y},\gamma) = \frac{n-k}{2}\log_2(\frac{\hat{\Phi}_{\hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y}|\boldsymbol{x})}}{n}) + \\ \frac{1}{2}\log_2\lambda + \frac{k}{2}\log_2\frac{\|\boldsymbol{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\gamma}\|_2^2}{n} - \\ \log_2\Gamma(\frac{n-k}{2}) - \log_2\Gamma(\frac{k}{2}) + \frac{n}{2}\log_2(n\pi). \end{aligned} \tag{22}$$

It is well-known that Lasso, Ridge, and Elastic net regression can have various values for their regularization parameters.

8) MDL for Polynomial Regression with Multi-Objective Goodness-of-Fit

Consider now set M as a set of polynomial regression models of degree $r \leq r'$. Denote $\hat{\boldsymbol{\theta}} = (\hat{a}_0, \ldots, \hat{a}_r)$ the set of coefficients in the polynomial of degree r. Since each $\hat{a}_s, s = 0, \ldots, r$ is a real number estimated from n data points, each $\hat{a}_s$ requires $\frac{1}{2}\log_2 n$ bits to encode, the same the code for degree r. Thus

$$L(\hat{\boldsymbol{\theta}}) = L(\hat{a}_0, \ldots, \hat{a}_r) = \frac{r+1}{2}\log_2 n + \frac{1}{2}\log_2 n = \frac{r+2}{2}\log_2 n. \tag{23}$$

The description of goodness-of-fit is

$$L(y|\hat{\boldsymbol{\theta}}) = \frac{n}{2}\log_2(\frac{RSS_r}{n}) \tag{24}$$

where $RSS_r = \sum_{i=1}^{n}(y_i - (\hat{a}_0 + \hat{a}_1 x_i + \cdots + (\hat{a}_r x_i^r))^2$. So the MDL for a polynomial of degree $r \leq r'$ is

$$MDL^{polREG}(\boldsymbol{y}, r) = \frac{r+2}{2}\log_2 n + \frac{n}{2}\log_2(\frac{RSS_r}{n}) \tag{25}$$

In our MDL, we propose to replace $\frac{RSS_r}{n}$ in (25) by a multi-objective criterion $\Phi_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})$, i.e.

$$MDL^{polREG}_{\Phi}(\boldsymbol{y}, r) = \frac{r+2}{2}\log_2 n + \frac{n}{2}\log_2(\frac{\hat{\Phi}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})}{n}). \tag{26}$$

9) MDL for a Feed-Forward Neural Network with Multi-Objective Goodness-of-Fit

We generally consider a feed-forward network (FF-NN) with $k \geq 1$ hidden layers, each having h_s hidden units, $s = 1, \ldots, k$ and m input and p output units. We propose a simple encoding for such models where the model description considers the encodings based on the encoding of the structure of the FF-NN and on the encoding of the learning part.

The encoding of the structure. The structure will be encoded as number of weights. In all hidden layers and in the output layer we considered ReLU activation function, and this is fixed for all FF-NN models.

The encoding of the learning part. We encode the learning part of the FF-NN models so that we encode the learning rate *lrt* of the Adam optimizer, the batch size *bs* and the number of Adam hyperparameters *Ahyp*. Denote the vector of all parameters defining a FF-NN by $\boldsymbol{\theta}$. We do not encode the values of $\boldsymbol{\theta}$ explicitely, but they are implicitely given by using Adam for their computation. Then

$$\begin{aligned} MDL^{FF-NN}(\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{2}\log_2(m\times h_1)+ \\ \frac{1}{2}\log_2(h_1\times h_2)+\cdots+\frac{1}{2}\log_2(h_k\times p) \\ +\frac{1}{2}\log_2(lrt)+\frac{1}{2}\log_2(bs)+\frac{1}{2}\log_2(Ahyp)+ \\ \frac{n}{2}\log_2(\frac{RSS_{FF-NN}}{n}) \end{aligned} \tag{27}$$

where RSS_{FF-NN} is the residual sum of squares on the output of $FF-NN$ and the values $\boldsymbol{y}$ and

$$\begin{aligned} MDL_{\Phi}^{FF-NN}(\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{2}\log_2(m\times h_1)+\frac{1}{2}\log_2(h_1\times h_2) \\ +\cdots+\frac{1}{2}\log_2(h_k\times p)+ \\ \frac{1}{2}\log_2(lrt)+\frac{1}{2}\log_2(bs)+\frac{1}{2}\log_2(Ahyp)+ \\ \frac{n}{2}\log_2(\frac{\hat{\Phi}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})}{n}). \end{aligned} \tag{28}$$

where $\Phi_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})$ is the multi-objective criterion applied on the output of $FF-NN$ and the values of $\boldsymbol{y}$.

D. ANALYTIC HIERARCHY PROCESS (AHP)

The Analytic Hierarchy Process (AHP) is a structured technique (Multiple Criteria Decision Analysis) for organizing and analyzing complex decisions [30]. It involves breaking down a problem into a hierarchy of subproblems that can be more easily comprehended and evaluated. The main steps in *AHP* are [31]:

- To decompose the decision problem into a hierarchy.
- To compare the elements at each hierarchy level to establish priorities.
- To synthesize these comparisons to determine weights for each element.

The consistency ratio (CR) [17] is calculated to ensure consistency in the comparisons:

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \tag{29}$$

where CI is the consistency index, and RI is the random index.

The consistency index (CI) measures the consistency of the pairwise comparisons. It is calculated as follows:

$$\mathrm{CI} = \frac{\lambda_{\max}-n}{n-1} \tag{30}$$

where $\lambda_{\max}$ is the largest eigenvalue of the comparison matrix, and n is the number of items being compared.

The random index (RI) is the average consistency index of a randomly generated pairwise comparison matrix. The value of the RI depends on the number of items being compared and is used as a benchmark to assess the acceptability of the calculated CI.

To implement *AHP* for model selection, we start by defining the criteria for model evaluation. For example, criteria such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Median Absolute Percentage Error (MdMAPE), Adjusted R-Squared (AdjR2), Prediction at 0.25 (Pred(0.25)), and Weighted Quantile Loss (wQL) can be used. Each model is evaluated based on these criteria through pairwise comparisons to determine their relative importance. The *AHP* process helps to synthesise these comparisons to assign a weight to each criterion, ultimately selecting the most suitable prediction model based on a comprehensive, structured evaluation. Each criterion has its own weight, which is set empirically or experimentally.

IV. EXPERIMENTS

This chapter outlines the experiments conducted using various regression model classes and two neural networks, namely multi-layered perceptron model classes to predict outcomes in the mentioned datasets. The experiments were divided into two main groups: Regression models and feed-forward neural networks. Each group utilized specific models' families, evaluated based on their performance with the corresponding datasets. All experiments were implemented using Python and libraries pandas, numpy, sklearn, tensorflow and intertools.

A. DATASETS

The datasets employed in this research are widely acknowledged and are publicly accessible. The historical data utilized in the study of software effort estimation is well established. In this work, we employ a Use Case Points (UCP, DS1) [32] and Glucose Level Prediction Dataset (GLP, DS2). The UCP-based dataset was first used in [32] and can be found in several following [33]–[35]. GLP dataset has been adopted from [36]. The Gross Domestic Dataset (GDP, DS3) has been used from [36], and finally, the Stock Market Prediction (STOCK, DS4) has been adopted from [37]. UCP and GLP datasets have been selected as they are are a mixture of real and synthetic samples. Expansion by synthetic samples is needed due to their small size. [36]. GDP a dataset is small in size and contains only real samples. Also in for STOCK dataset only real samples has been used, but there is 10,900 samples.

1) DS1 - Use Case Points Dataset

Use Case Points Dataset is based on ucp_71, which was first used in [32] and can be found in several following papers [33]–[35]. After removing outlier based on the interquartile range approach (*IQR*) [38], the UCP dataset contains 4,912 samples, with characteristics described in Table 1. Unadjusted Actor Weight (UAW) measures the complexity of actors interacting with a system. Unadjusted Use Case Weight (UUCW) assesses the complexity of use cases. Technical Complexity Factor (TCF) evaluates the technical aspects that affect a project's complexity. Environmental Complexity Factor (ECF) considers environmental factors impacting the project. Effort refers to the total amount of work required to complete a project, typically measured in person-hours or person-months. These variables are used to estimate and manage the scope and resources needed for software development projects.

TABLE 1: UCP Dataset Characteristics

Variable	Type	Mean	Median	Min	Max
UAW	indep.	10.51	6.00	8.00	19.00
UUCW	indep.	388.97	355.00	250.00	610.00
TCF	indep.	0.92	0.94	1.11	0.71
ECF	indep.	0.87	0.89	0.51	1.08
Effort	dep.	6,560.00	6,406.00	5,775.00	7,970.00

2) DS2 - Glucose Level Prediction Dataset

Selected GLP dataset [36] initially consists of 16,979 samples; after applying *IQR* cleaning [38], 15,942 samples have been prepared for further use. The dataset represents ten variables that describe various physiological and metabolic parameters. GLP dataset characteristics are in Table 2. The dependent variable is Blood Glucose Level (BGL), measured in milligrams per deciliter (mg/dL). The independent variable (predictor) is age (AGE), which records the age of individuals in years. Blood pressure readings are split into Diastolic Blood Pressure (DBP) and Systolic Blood Pressure (SDP), measured in millimetres of mercury (mmHg). Heart Rate (HR), expressed in beats per minute (bpm), reflects cardiovascular health. Body Temperature (TE), recorded in degrees Fahrenheit, can indicate fever or hypothermia. Next is a blood oxygen saturation (SPO2) as a percentage, highlighting respiratory efficiency. The dataset also includes categorical variables for sweating (SWE) and shivering (SHI), capturing the presence of these symptoms. Both evaluate yes/no value. Finally, the diabetic/non-diabetic (D/N) variable categorises individuals based on their diabetic status.

3) DS3 - Federal Reserve Bank

In this dataset, gross domestic product and inflation data from [39] with their basic descriptive statistics for the key economic indicators are presented in Table 3. In total, there are 97 samples (94 past quarters). The gross domestic product (GDP), in dollars has an average value of 16,846.501 with a median of 15,955.532, ranging from a minimum of 10,002.179 to a maximum of 27,956.998. Inflation, represented as an index, averages 225.936 with a median of 227.296, fluctuating between 169.300 and 307.531. The Interest Rate shows a mean of 1.786%, a median of 1.080%, and spans from 0.050% to 6.540%.

TABLE 2: GLP Dataset Characteristics

Variable	Type	Mean	Median	Min	Max
AGE	indep.	30.98	14.00	9.00	77.00
DBP	indep.	77.17	76.00	60.00	95.00
SBP	indep.	118.18	119.00	95.11	145.00
TE	indep.	97.35	97.32	96.00	98.08
SPO2	indep.	97.38	98.00	93.00	99.00
SWE	indep.		categorical		
SHI	indep.		categorical		
D/N	indep.		categorical		
BGL	dep.	86.72	82.00	50.00	129.00

The Unemployment Rate has a mean of 5.805% and a median of 5.250%, with values ranging from 3.40% to 14.80%. Consumer Sentiment averages at 83.580, with a median of 86.450, and varies from a low of 51.500 to a high of 112.000. Industrial Production has an average of 96.974 and a median of 98.580, with a minimum of 84.681 and a maximum of 103.929.

Money Supply, another key economic indicator, has an average of 10,893.520 and a median of 9,647.700, with values ranging from 4,666.200 to 21,722.300. Finally, Personal Income averages 14,248.143 with a median of 13,519.000 and spans from a minimum of 8,348.000 to a maximum of 23,189.400.

4) DS4 - Stock Price

The fourth dataset is adapted from [37], and the dataset characteristics of stock prices are summarized Table 4. The table presents basic descriptive statistics of five key variables: Date, Open, High, Low, and Close prices. Each variable's type is identified as Independent or Dependent, with Close being the dependent variable. There are 10,900 samples in total.

The Date variable, marked as Independent, is represented in an ordinal format, signifying the timestamp of each record. The market prices Open, High, and 'Low' are classified as independent variables. The Open price has a mean value of 82.350993, a median of 46.948750, a minimum value of 6.870357, and a maximum value of 453.070007. The 'High' price shows a mean of 83.178018, a median of 47.371250, a minimum of 7.000000, and a maximum of 456.170013. Similarly, the 'Low' price has a mean of 81.526316, a median of 46.549999, a minimum of 6.794643, and a maximum of 451.769989.

The Close price, being the dependent variable, has a mean of 82.388479, a median of 47.000000, a minimum of 6.858929, and a maximum of 452.850006.

TABLE 3: GDP Dataset Characteristics

Variable	Type	Mean	Median	Min	Max
GDP	dep.	16846.501	15955.532	10002.179	27956.998
Inflation	dep.	225.935	227.296	169.300	307.531
Interest_Rate	indep.	1.786	1.080	0.050	6.540
Unemployment_Rate	indep.	5.805	5.250	3.400	14.800
Consumer_Sentiment	indep.	83.580	86.450	51.500	112.000
Industrial_Production	indep.	96.974	98.580	84.681	103.929
Money_Supply	indep.	10893.520	9647.700	4666.200	21722.300
Personal_Income	indep.	14248.143	13519.000	8348.000	23189.400

TABLE 4: Stock Price Dataset Characteristics

Variable	Type	Mean	Median	Min	Max
Date	indep.	ordinal representation of timestamp			
Open	indep.	82.350993	46.948750	6.870357	453.070007
High	indep.	83.178018	47.371250	7.000000	456.170013
Low	indep.	81.526316	46.549999	6.794643	451.769989
Close	dep.	82.388479	47.000000	6.858929	452.850006

B. REGRESSION MODEL CLASSES

Multiple Linear Regression Model Class

The linear regression models tested in this work utilized various configurations. The multiple linear regression (LR) model is tested for all combinations of available predictors.

The Ridge LR, Lasso and Elastic net model classes were evaluated with different values for the regularization parameter alpha (0.1, 1.0, 10.0, 100.0) and combination of predictors.

Polynomial regression models were constructed to capture non-linear relationships within the datasets. These models were tested with two different polynomial degrees and their respective hyperparameter parameters, i.e. vectors of polynomial coefficients.

For quadratic regression models (i.e polynomial regression up to degree, the hyperparameters included setting the maximum polynomial degree up to 2. Similarly, cubic polREG used the same configuration up to degree 3

These configurations and the selected best parameters allowed for a comprehensive evaluation of the data's linear and non-linear relationships, ensuring that the models could capture the underlying patterns effectively.

C. FEED-FORWARD NEURAL NETWORK MODEL CLASS

Feed-forward neural network models (FF-NN) was selected to address the prediction task in this experiment. The model was designed sequentially, starting with an input layer matching our feature set's dimensionality. Multiple configurations of hidden layers were tested, comprising varying numbers of neurons and dropout rates to prevent overfitting.

The architecture of FF-NN I consists of one input layer, two hidden layers, and one output layer. The hidden layers have 64 and 32 neurons, respectively. This setup is often used in practice as it is simple yet powerful enough to capture complex patterns in data without overfitting. A dropout rate of 0.3 is used, effectively preventing overfitting by randomly deactivating 30% of the neurons during training. Dropout is a widely accepted regularisation technique that helps improve the generalisation of neural networks [40].

A learning rate of 0.0001 is chosen for the training details, allowing the model to converge smoothly to a minimum. A lower learning rate helps fine-tune the weights more accurately [41]. The Adam Optimiser was selected for its efficiency and capability to adapt the learning rate during training. Adam combines the advantages of both the AdaGrad and RMSProp algorithms [42]. The model is trained for 200 epochs with a batch size 16, ensuring that the model sees enough data instances for robust training while maintaining computational efficiency. To prevent overfitting and ensure optimal performance, early stopping is implemented. The training stops if the validation loss does not improve for 10 consecutive epochs, with the best model weights restored [43].

The architecture of FF-NN II includes one input layer, three hidden layers, and one output layer with neurons arranged as [128, 64, 32]. The configuration is summarized in Table 5. Each hidden layer utilised the ReLU activation function, which is known for its efficiency in training deep networks by mitigating the vanishing gradient problem. The dropout layers were strategically inserted after each dense layer to regularise the model and improve generalisation by randomly setting a fraction of input units to zero during the training phase. The output layer consisted of a single neuron with a ReLU activation function, which was suitable for our regression objective.

TABLE 5: Multi-layered perceptron configuration

Model	**Configuration**	**Training Details**
FF-NN I	1 Input, 2 Hidden, 1 Output Layer Layers: [64, 32] Dropout: 0.3 Learning Rate: 0.0001 Optimizer: Adam	Epochs: 200 Batch Size: 16 Early Stopping: monitor='val_loss', patience=10, restore_best_weights=True
FF-NN II	1 Input, 3 Hidden, 1 Output Layer Layers: [128, 64, 32] Dropout: 0.2 Learning Rate: 0.001 Optimizer: Adam	Epochs: 200 Batch Size: 16 Early Stopping: monitor='val_loss', patience=10, restore_best_weights=True

D. LIMITATIONS AND INTERPRETATION OF THE RESULTS

This subsection discusses the limitations of our work and cautions us in interpreting the results and conclusions.

1) Limitations

- Datasets vary in size and complexity; smaller datasets like GDP may limit generalization. We added synthetic samples and used holdout validation to enhance robustness.
- Integrating MDL and MCDA with FF-NN models requires significant resources and can hinder interpretability. We used early stopping (10 epochs without validation loss improvement) and documented model details to address this.
- MCDA's weighting of criteria introduces subjectivity. We minimized this with established methods and expert input, suggesting future studies explore objective weighting.
- Neural network performance relies on hyperparameter tuning. We used grid and random search with holdout validation to ensure robustness.

2) Caution in Interpretation

- FF-NN models risk overfitting; we used holdout validation and early stopping to address this.
- Results may not generalize beyond the datasets used. We included diverse datasets and suggested future studies to test the MDL-MCDA framework on broader datasets.
- Model selection may introduce bias. We tested various models and recommend including emerging techniques in future work.
- Data inaccuracies can affect performance. We ensured data quality through trusted sources, outlier detection, and handling missing values.

In summary, while our work demonstrates the potential benefits of integrating MDL and MCDA for model selection, the above issues highlight areas for caution.

V. RESULTS AND DISCUSSION

The results of all regression and FF-NN models are compared and evaluated for all four tested datasets - problem domains. Each model is assessed based on the *AHP*, MDL_{AHP} and MDL_{RSS}. *AHP* score is constructed using measures described in Section III-B. For regression models and FF-NN models, results tables in Sections V-A and V-B contain the best variants where, where models where MDL_{AHP} reach its minimal value for the highest value of *AHP*. Detailed results and discussion are available in Appendix

A. REGRESSION MODELS

The performance of regression models using the UCP dataset is evaluated based on *AHP*, MDL_{AHP}, and MDL_{RSS} is presented in In Table 6. The best regression model for the UCP dataset is the Polynomial_3 model with UAW, UUCW, and ECF predictors, achieving an *AHP* Score of 1.00, MDL_{AHP} of -12196.06, and MDL_{RSS} of 52135.34.

For the GLP dataset the polynomial regression of degree 3 using AGE, DBP, SBP, TE, SPO2, SWE, HR, SHI, and DN predictors achieves the highest *AHP* Score of 1.00, with MDL_{AHP} of -42044.46 and MDL_{RSS} of 113114.61. In the GDP dataset, the best regression model is the Polynomial Regression of Degree 3 using Interest Rate, Industrial Production, Money Supply, and Personal Income predictors, with an *AHP* Score of 1.00, MDL_{AHP} of 986.16, and MDL_{RSS} of 1858.92. For the STOCK dataset, the Polynomial Regression of Degree 3 with Open, High, and Low predictors achieves an *AHP* Score of 1.00, MDL_{AHP} of -32230.31, and MDL_{RSS} of 22619.67.

B. FEED-FORWARD NEURAL NETWORK MODELS

Table 7 presents the performance of FF-NN models on the UCP dataset. The best model is FF-NN II with UAW, UUCW, TCF, and ECF predictors, achieving an *AHP* Score of 1.00, MDL_{AHP} of -30093.69, and MDL_{RSS} of 35846.81.

For the GLP dataset FF-NN II with AGE, DBP, SBP, TE, SPO2, HR, SHI, and DN predictors achieves the highest *AHP* Score of 0.99, with MDL_{AHP} of -111365.58 and MDL_{RSS} of 42619.76. In the GDP dataset, the best FF-NN model is FF-NN II with industrial production + personal income predictors, achieving an *AHP* Score of 1.00, MDL_{AHP} of -294.06, and MDL_{RSS} of 849.81. For the STOCK dataset FF-NN II with Low as the predictor achieves the highest *AHP* Score of 1.00, MDL_{AHP} of -73376.47, and MDL_{RSS} of -11413.85. Including market price predictors consistently improved model performance across the various regression techniques, as evidenced by higher *AHP* Scores and more favourable MDL_{AHP} and MDL_{RSS} values compared to using the Date predictor alone.

C. DISCUSSION

The final comparison is displayed in the two plots for FF-NN models (Figure 2) and regression models (Figure 3). The performance of three criteria: *AHP*, and MDL_{RSS} across four datasets: GLP, STOCK, UCP, and GDP for FF-NN models and regression models are shown. The *AHP* criterion is depicted in blue; the MDL_{AHP} criterion is represented in red; and the MDL_{RSS} criterion is shown in green. The comparison between regression models and feed-forward neural networks FF-NN models reveals that FF-NN models generally outperform regression models across all datasets in terms of MDL_{AHP}.

For instance, in the GDP dataset, the FF-NN II model with Industrial Production and Personal Income predictors achieved an MDL_{AHP} of -294.06 compared to the Polynomial Regression's MDL_{AHP} of 986.16. Similarly, for the GLP dataset, FF-NN II achieved MDL_{AHP} of -111365.58, outperforming the Polynomial Regression's MDL_{AHP} of -42044.46. The UCP and STOCK datasets show similar trends, with FF-NN models consistently having better MDL_{AHP} values than regression models.

Across all datasets, MDL_{AHP} consistently provides better performance measure than MDL_{RSS}. For example, in the STOCK dataset, FF-NN II with Low as the predictor

TABLE 6: Performance of Regression Models across Datasets

Dataset	Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
UCP	Polynomial_3	UAW+UUCW+ECF	1.00	-12196.06	52135.34
GLP	Polynomial_3	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	1.00	-42044.46	113114.61
GDP	Polynomial_3	Interest Rate+Industrial Production+Money Supply+Personal Income	1.00	986.16	1858.92
STOCK	Polynomial_3	Open+High+Low	1.00	-32230.31	22619.67

TABLE 7: Performance of FF-NN Models across Datasets

Dataset	Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
UCP	FF-NN II	UAW,UUCW,TCF,ECF	1.00	-30093.69	35846.81
GLP	FF-NN II	AGE,DBP,SBP,TE,SPO2,HR,SHI,DN	0.99	-111365.58	42619.76
GDP	FF-NN II	Industr. Production, Personal Income	1.00	-294.06	849.81
STOCK	FF-NN II	Low	1.00	-73376.47	-11413.85

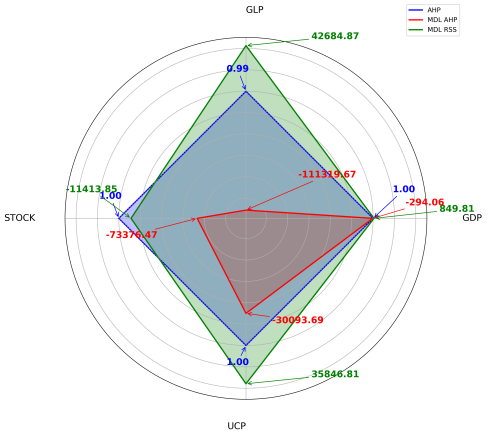

FIGURE 2: Comparison of Selected FF-NN Models.

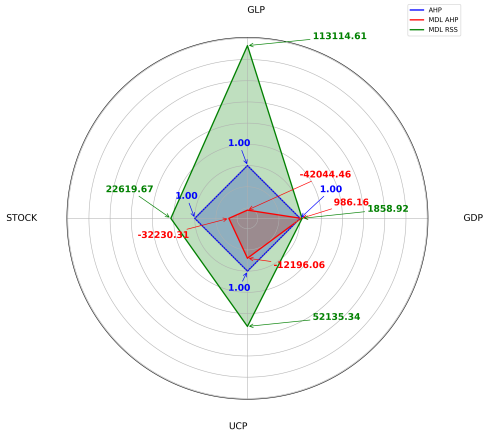

FIGURE 3: Comparison of Selected Regression Models.

achieved an MDL_{AHP} of -73376.47 compared to its MDL_{RSS} of -11413.85, indicating a significant contrast. This trend is observed across all datasets, suggesting that MDL_{AHP} better captures the trade-offs and multi-criteria evaluations inherent in complex model selection compared to MDL_{RSS}.

However, while FF-NNs provide significant advantages due to their non-linearity and depth, they also come with higher computational costs and complexity, which can be a disadvantage regarding interpretability and ease of implementation. However, it must consider that MDL_{AHP} may also involve more subjective judgment in determining weights for different criteria, which can introduce bias.

VI. CONCLUSION

This work aimed to investigate the integration of the Minimum Description Length (MDL) principle with Multi-Criteria Decision Analysis (MCDA) to enhance model selection for predictive tasks.

In conclusion, this study introduces a novel aproach that integrates the MDL principle with the AHP, a form of multi-criteria decision analysis, to address critical challenges in predictive modeling and model selection. The primary motivation for combining MDL and AHP arises from the limitations of traditional MDL approaches, which typically rely on the Residual Sum of Squares (RSS) as a sole measure of goodness-of-fit. By integrating MDL with AHP, we extend the model evaluation process beyond a single error metric, allowing for a more robust and nuanced assessment that accounts for multiple evaluation criteria. Through a comprehensive analysis, we addressed and contributed to these research questions:

- **RQ1: How do the Minimum Description Length (MDL) and Multi-Criteria Decision Analysis (MCDA) integration affect predictive model selection?**
  The integration of MDL and MCDA has significantly impacted model selection. The results across various datasets demonstrated that the combined MDL-MCDA approach consistently outperformed traditional model selection methods based solely on MDL or MCDA. Incorporating MCDA allowed for a balanced considera-

tion of multiple evaluation measures, addressing the conflicting criteria inherent in predictive modelling. This integration ensured a robust selection process, improving the overall accuracy and reliability of the predictive models.

- **RQ2: What advantages does the MDL-MCDA have compared to the MDL-RSS in predictive model selection?**
  When comparing the MDL-MCDA method to the traditional MDL-RSS approach, the findings indicated a clear advantage of MDL-MCDA. The MDL-MCDA method provided a more nuanced evaluation by incorporating multiple criteria and demonstrated superior performance in selecting models that generalised well to new data. The MDL-RSS approach, while effective in some scenarios, often needed to improve in balancing the complexity and fit of the models, leading to suboptimal selections in datasets.

This work opens several directions for future research. Firstly, further integration of multi-criteria methods can be investigated. Exploring other multi-criteria decision-making methods could provide additional insights and improve model selection processes.

Secondly, an application to different domains should be explored. Applying the MDL-MCDA framework to other domains and datasets will help to further validate its versatility and robustness across various predictive modelling tasks.

Thirdly, developing automated methods for determining the weights of evaluation criteria in MCDA could reduce subjective bias and enhance the objectivity of the model selection process.

In conclusion, the integration of MDL and MCDA presents a promising approach to model selection for predictive tasks, offering a balanced and comprehensive framework that addresses the inherent complexities and conflicting criteria of predictive modelling.

REFERENCES

[1] F. Petropoulos, N. Kourentzes, K. Nikolopoulos, and E. Siemsen. Judgmental selection of forecasting models. *Journal of Operations Management*, 60:34–46, 2018.

[2] Jerome H Friedman. An overview of predictive learning and function approximation. *From statistics to neural networks: Theory and pattern recognition applications*, pages 1–61, 1994.

[3] Aurélie Boisbunon. *Model selection: a decision-theoretic approach*. Thesis, Université de Rouen, 2013.

[4] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[5] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[6] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[7] M. A. Villegas, D. J. Pedregal, and J. R. Trapero. A support vector machine for model selection in demand forecasting applications. *Computers & Industrial Engineering*, 121:1–7, 2018.

[8] A. A. Ghobbar and C. H. Friend. Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers & Operations Research*, 30(14):2097–2114, 2003.

[9] Chi-Ok Oh and Bernard J. Morzuch. Evaluating time-series models to forecast the demand for tourism in singapore:comparing within-sample and postsample results. *Journal of Travel Research*, 43(4):404–413, 2005.

[10] J. W. Taylor and P. E. McSharry. Short-term load forecasting methods: An evaluation based on european data. *Ieee Transactions on Power Systems*, 22(4):2213–2219, 2007.

[11] W. W. Han, X. Wang, F. Petropoulos, and J. Wang. Brain imaging and forecasting: Insights from judgmental model selection. *Omega-International Journal of Management Science*, 87:1–9, 2019.

[12] Abbas Mardani, Ahmad Jusoh, Khalil Md Nor, Zainab Khalifah, Norhayati Mohmad Zakwan, and Alireza Valipour. Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014, 01 2015.

[13] Ian Durbach and Theodor J. Stewart. Modeling uncertainty in multi-criteria decision analysis, 11 2012.

[14] Yvonne Badulescu, Ari-Pekka Hameri, and Naoufel Cheikhrouhou. Evaluating demand forecasting models using multi-criteria decision-making approach, 10 2021.

[15] Ashu Bansal, Brijesh Kumar, and Rakesh Garg. Multi-criteria decision making approach for the selection of software effort estimation model, 01 2017.

[16] Luís G. Vargas and John J. Dougherty. The analytic hierarchy process and multicriterion decision making, 01 1982.

[17] T. L. Saaty. How to make a decision - the analytic hierarchy process. *European Journal of Operational Research*, 48(1):9–26, 1990.

[18] Federica Sarro, Alessio Petrozziello, and Mark Harman. Multi-objective software effort estimation, 05 2016.

[19] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[20] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[21] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific, 1998.

[22] Jorma Rissanen. *Information and Complexity in Statistical Modeling*. Springer Science & Business Media, 2007.

[23] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[24] Peter Grünwald and Teemu Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, 2020.

[25] Thomas CM Lee. An introduction to coding theory and the two-part minimum description length principle. *International statistical review*, 69(2):169–183, 2001.

[26] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[27] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[28] Ciprian Doru Giurcăneanu, Seyed Alireza Razavi, and Antti Liski. Variable selection in linear regression: Several approaches based on normalized maximum likelihood. *Signal Processing*, 91(8):1671–1692, 2011.

[29] Jorma Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543, 2000.

[30] Thomas L Saaty. What is the analytic hierarchy process? In *Mathematical models for decision support*, pages 109–121. Springer, 1988.

[31] Madjid Tavana, Mehdi Soltanifar, and Francisco J Santos-Arteaga. Analytical hierarchy process: revolution and evolution. *Annals of operations research*, 326(2):879–907, 2023.

[32] Radek Silhavy, Petr Silhavy, and Zdenka Prokopova. Analysis and selection of a regression model for the use case points method using a stepwise approach. *Journal of Systems and Software*, 125:1–14, 2017.

[33] Z. Prokopova, R. Silhavy, and P. Silhavy. The effects of clustering to software size estimation for the use case points methods. In *Advances in Intelligent Systems and Computing*, volume 575, pages 479–490. 2017. Export Date: 29 May 2017.

[34] P. Silhavy, R. Silhavy, and Z. Prokopova. Evaluation of data clustering for stepwise linear regression on use case points estimation. In *Advances in Intelligent Systems and Computing*, volume 575, pages 491–496. 2017. Export Date: 29 May 2017.

[35] Radek Silhavy, Petr Silhavy, and Zdenka Prokopova. Evaluating subset selection methods for use case points estimation. *Information and Software Technology*, 97:1–9, 2018.

[36] Deepali Javale and Sharmishta Desai. Dataset for people for their blood glucose level with their superficial body feature readings., 2021.

[37] Yahoo Finance. Stock market data. https://finance.yahoo.com. Accessed: 2024-07-11.

[38] Ch Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, and Ashish Ghosh. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6:100164, 2023.
[39] Federal Reserve Bank of St. Louis. Gross domestic product, 2024. Accessed: 2024-07-10.
[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.
[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[43] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

PETR SILHAVY is an Associate Professor with the Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic. He is a Senior Research and Associate Professor of System Engineering and Informatics with a demonstrated history of working in research and higher education. He has expertise as a CTO and a Software Developer in database programming, database design, data management, and data science. Petr received his PhD in Engineering Informatics (2009) at the Faculty of Applied Informatics, Tomas Bata University in Zlin. His research interests are prediction and empirical methods for software engineering.

KATEŘINA HLAVÁČKOVÁ-SCHINDLER is a senior scientist in the Data Mining and Machine Learning research group at the University of Vienna, Vienna, Austria. She recieved her MSc summa cum laude in mathematics in Charles University in Prague, Czech Republic, PhD in computer science in the Czech Academy of Sciences and habilitation (=Privatdoz) in University of Vienna. She has more than 80 publications mostly on causal inference and causal discovery, machine learning and artificial neural networks.

RADEK SILHAVY is an Associate Professor and a Senior Researcher with the Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic. He is an Associate Professor of System Engineering and Informatics with a demonstrated history of working in research, higher education, project management, and software analysis. He is also involved in academic publishing as the editor-in-chief, editor, or reviewer. Radek received his PhD in Engineering Informatics (2009) at the Faculty of Applied Informatics, Tomas Bata University in Zlin. His research interests are predictive analytics for software engineering, empirical methods in software engineering, or prediction models focused on cost, size, and effort estimations in system/software engineering.

APPENDIX. DETAILED RESULTS TO REGRESSION MODELS AND NEURAL NETWORK MODELS

All regression and FF-NN models' results are compared and evaluated for all four tested datasets - problem domains. Each model is assessed based on the *AHP*, MDL_{AHP} and MDL_{RSS}. *AHP* score is constructed using measures described in the Section III-B. For regression models in the tables included, the best variants according to one of the criteria - a model where *AHP* is maximal then a model where MDL_{AHP} or MDL_{RSS} is minimal.

A. REGRESSION MODELS

In Table 8 the Analytical Hierarchy Process (AHP) score, the Minimum Description Length based on *AHP* (MDL_{AHP}), and the Minimum Description Length based on Residual Sum of Squares (MDL_{RSS}) is presented.

The predictors for each model are also specified, providing insight into the input variables considered for the regression analysis.

The ElasticNet model with the UAW predictor achieved an *AHP* Score of 0.66, an MDL_{AHP} of -14096.66, and an MDL_{RSS} of 56886.55. This indicates a moderate *AHP* score but relatively high MDL values, suggesting potential overfitting.

Models incorporating the UAW, UUCW, and ECF predictors generally exhibited higher performance measures. The Lasso, LinearRegression, Polynomial_2, Polynomial_3, and Ridge models with these predictors achieved an *AHP* Score of 0.70 or higher. Notably, the Polynomial_3 model achieved the highest *AHP* Score of 1.00, with an MDL_{AHP} of -12196.06 and an MDL_{RSS} of 52135.34, indicating superior performance in capturing the underlying patterns in the dataset.

The models using TCF as the sole predictor (ElasticNet, Lasso, LinearRegression, Polynomial_2, Polynomial_3, and Ridge) uniformly resulted in an *AHP* Score of 0.59. However, these models exhibited varying MDL_{AHP} and MDL_{RSS} values, with Polynomial_3 again showing a relatively better performance with an MDL_{AHP} of -14069.16 and an MDL_{RSS} of 58082.99.

When examining the combination of UAW and ECF predictors, the ElasticNet model achieved an *AHP* Score of 0.66, an MDL_{AHP} of -14097.32, and an MDL_{RSS} of 56885.56. This result is comparable to the model's performance with UAW alone, suggesting that ECF may not significantly enhance the model's predictive capability in this context.

In summary, for UCP, the Polynomial_3 model with UAW, UUCW, and ECF predictors consistently outperforms other models, achieving the highest *AHP* Score and the lowest MDL values. The results underscore the importance of selecting appropriate predictors and model complexity to enhance the regression model's performance.

Table 9 presents *AHP*, MDL_{AHP} and MDL_{RSS} for GLP Dataset. The model with the highest *AHP* Score is the Polynomial Regression of Degree 3, with a score of 1.00, using the predictors AGE, DBP, SBP, TE, SPO2, SWE, HR, SHI, and DN. This indicates that this model performs best regarding *AHP* among the models.

When considering MDL_{AHP}, the Polynomial Regression of Degree 3 again shows the best performance with a value of -42044.46. However, the Linear Regression and Ridge Regression models, both using the predictors AGE, DBP, SBP, TE, SPO2, SWE, HR, SHI, and DN, also show strong performance with MDL_{AHP} values of -51369.50.

For MDL_{RSS}, the Linear Regression model with the predictors AGE, DBP, SBP, TE, SPO2, SWE, HR, SHI, and DN performs best with the lowest value of 106125.93. The Ridge Regression model using the same predictors shows an identical performance in MDL_{RSS} with a value of 106125.92.

Overall, the Polynomial Regression of Degree 3 with the predictors AGE, DBP, SBP, TE, SPO2, SWE, HR, SHI, and DN is the best model when considering both *AHP* and MDL_{AHP}. The Linear Regression and Ridge Regression models are the best when MDL_{RSS} is considered. Therefore, if all three criteria are considered, the Polynomial Regression of Degree 3 emerges as the most optimal model due to its superior performance in two out of three measures.

Table 10 presents *AHP*, MDL_{AHP} and MDL_{RSS} for GDP Dataset. The model with the highest *AHP* Score is the Polynomial Regression of Degree 3 using the predictors of Interest Rate, Industrial Production, Money Supply, and Personal Income, achieving an *AHP* Score of 1.00. This indicates that this model performs the best in *AHP* among all evaluated models.

In terms of MDL_{AHP}, the Polynomial Regression of Degree 2 using the predictors' Unemployment Rate, Consumer Sentiment, Industrial Production, Money Supply, and Personal Income shows the highest value of 1419.81, suggesting it is the best model based on this criterion. However, the Polynomial Regression of Degree 3 with predictors Interest Rate and Personal Income also shows strong performance with a MDL_{AHP} value of 986.16.

Regarding MDL_{RSS}, the Linear Regression model with the predictors of Industrial Production and Personal Income performs best with the lowest value of 1066.64, indicating it has the least residual sum of squares. The Lasso and ElasticNet models using similar predictors also demonstrate competitive performance in this measure.

The Polynomial Regression of Degree 3 with the predictors of Interest Rate, Industrial Production, Money Supply, and Personal Income is the best model when considering the *AHP* Score. The Polynomial Regression of the Degree 2 model stands out in terms of MDL_{AHP}, and the Linear Regression model excels in MDL_{RSS}. Thus, if all three criteria are taken into account, the Polynomial Regression of Degree 3 is the most optimal model due to its superior performance in *AHP* Score and competitive performance in MDL_{AHP} and MDL_{RSS}

Table 11 presents *AHP*, MDL_{AHP} and MDL_{RSS} for STOCK Dataset. The models tested include ElasticNet, Lasso, Linear Regression, Polynomial Regression (of degree 2 and 3), and Ridge Regression. Each model was evaluated with different predictors, specifically Open+High+Low and Date.

The ElasticNet model, when using Open+High+Low as predictors, achieved an *AHP* Score of 0.52, an MDL_{AHP} of -37733.67, and an MDL_{RSS} of 67881.56. When using Date as

TABLE 8: *AHP*, MDL_{AHP} and MDL_{RSS} for Regression Models with UCP dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
ElasticNet	UAW	0.66	-14096.66	56886.55
Lasso	UAW+UUCW+ECF	0.70	-13880.71	56475.11
LinearRegression	UAW+UUCW+ECF	0.70	-13879.19	56474.57
Polynomial_2	UAW+UUCW+ECF	0.90	-12830.80	53385.08
Polynomial_3	UAW+UUCW+ECF	1.00	-12196.06	52135.34
Ridge	UAW+UUCW+ECF	0.70	-13879.35	56474.56
ElasticNet	TCF+ECF	0.59	-14485.98	57663.44
Lasso	TCF	0.59	-14482.79	57668.99
LinearRegression	TCF	0.59	-14482.51	57668.63
Polynomial_2	TCF	0.59	-14324.29	57826.49
Polynomial_3	TCF	0.59	-14069.16	58082.99
Ridge	TCF	0.59	-14482.52	57668.63
ElasticNet	UAW+ECF	0.66	-14097.32	56885.56

TABLE 9: *AHP*, MDL_{AHP} and MDL_{RSS} for Regression Models with GLP dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
ElasticNet	DBP+SBP+TE+HR+SHI+DN	0.89	-51688.56	106567.37
ElasticNet	AGE+DN	0.53	-57605.95	111464.48
ElasticNet	AGE+DBP+SBP+TE+SPO2+HR+SHI+DN	0.89	-51712.63	106555.24
Lasso	DBP+SBP+TE+SPO2+HR+SHI+DN	0.91	-51505.11	106308.21
Lasso	AGE+DN	0.53	-57593.60	111478.94
Lasso	DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	0.91	-51506.30	106308.21
LinearRegression	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	0.92	-51369.50	106125.93
LinearRegression	AGE	0.53	-57652.15	111455.88
Polynomial_2	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	0.97	-48430.90	107477.24
Polynomial_2	DN	0.54	-55271.50	113769.02
Polynomial_3	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	1.00	-42044.46	113114.61
Polynomial_3	DN	0.54	-49213.66	119826.67
Ridge	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	0.92	-51369.50	106125.92
Ridge	AGE	0.53	-57652.15	111455.88

TABLE 10: Regression Results for *AHP* and MDL on GDP

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
ElasticNet	Interest_Rate+Industrial_Production+Money_Supply+Personal_Income	0.70	0.19	1207.44
ElasticNet	Consumer_Sentiment	0.21	-84.13	1407.13
Lasso	Industrial_Production+Personal_Income	0.80	10.66	1066.71
Lasso	Interest_Rate+Consumer_Sentiment	0.19	-88.20	1409.81
Lasso	Interest_Rate+Industrial_Production+Personal_Income	0.80	10.55	1066.69
LinearRegression	Industrial_Production+Personal_Income	0.80	10.66	1066.65
LinearRegression	Interest_Rate+Consumer_Sentiment	0.19	-88.21	1409.82
LinearRegression	Interest_Rate+Industrial_Production+Personal_Income	0.80	10.55	1066.64
Polynomial_2	Unemployment_Rate+Consumer_Sentiment+Industrial_Production+Money_Supply+Personal_Income	0.91	1419.81	2360.46
Polynomial_2	Consumer_Sentiment	0.20	1313.57	2811.79
Polynomial_2	Interest_Rate+Personal_Income	0.91	1419.37	2348.95
Polynomial_3	Interest_Rate+Industrial_Production+Money_Supply+Personal_Income	1.00	986.16	1858.92
Polynomial_3	Consumer_Sentiment	0.21	877.40	2368.80
Ridge	Interest_Rate+Industrial_Production+Personal_Income	0.79	9.87	1074.49
Ridge	Interest_Rate+Consumer_Sentiment	0.17	-97.24	1409.57

the predictor, the *AHP* Score dropped to 0.04, while MDL_{AHP} and MDL_{RSS} were -57585.61 and 98004.58, respectively. The Lasso model with Low as the predictor had an *AHP* Score of 0.80, MDL_{AHP} of -34271.85, and MDL_{RSS} of 32672.06, whereas with Date as the predictor, the *AHP* Score was 0.04, MDL_{AHP} was -57584.76, and MDL_{RSS} was 98005.44.

Linear Regression using Open+High+Low predictors achieved perfect *AHP* Scores of 1.00 with MDL_{AHP} of -32535.83 and MDL_{RSS} of 22562.16, while with Date as the predictor, the *AHP* Score was 0.04, MDL_{AHP} was -57584.69, and MDL_{RSS} was 98005.51. Polynomial Regression models (degrees 2 and 3) with Open+High+Low predictors also achieved perfect *AHP* Scores of 1.00, with MDL_{AHP} and MDL_{RSS} values of -32444.69 and 22851.12 for degree 2, and -32230.31 and 22619.67 for degree 3. Including the Date predictor alongside Open+High+Low in Polynomial Regression (degree 3) yielded similar results.

The Ridge Regression model with Open+High+Low pre-

TABLE 11: *AHP*, MDL_{AHP} and MDL_{RSS} for Regression Models with STOCK dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
ElasticNet	Open+High+Low	0.52	-37733.67	67881.56
ElasticNet	Date	0.04	-57585.61	98004.58
Lasso	Low	0.80	-34271.85	32672.06
Lasso	Date	0.04	-57584.76	98005.44
LinearRegression	Open+High+Low	1.00	-32535.83	22562.16
LinearRegression	Date	0.04	-57584.69	98005.51
Polynomial_2	Open+High+Low	1.00	-32444.69	22851.12
Polynomial_2	Date	0.04	-57474.61	98115.58
Polynomial_3	Open+High+Low	1.00	-32230.31	22619.67
Polynomial_3	Date	0.04	-57285.34	98304.85
Polynomial_3	Date+Open+High+Low	1.00	-32230.31	22619.67
Ridge	Open+High+Low	0.93	-33062.29	25693.66
Ridge	Date	0.04	-57584.69	98005.51

dictors had an *AHP* Score of 0.93, MDL_{AHP} of -33062.29, and MDL_{RSS} of 25693.66. Using Date as the sole predictor resulted in an *AHP* Score of 0.04, MDL_{AHP} of -57584.69, and MDL_{RSS} of 98005.51.

B. FEED-FORWARD NEURAL NETWORKS - MULTI-LAYER PERCEPTRON

Table 12 summarizes the performance of FF-NN models on the UCP dataset, evaluated through the *AHP*, MDL_{AHP}, and MDL_{RSS}. Two models, FF-NN I and FF-NN II, are compared using different combinations of predictors.

The performance of FF-NN models on the UCP dataset was compared across various configurations of predictors. The results are summarised in Table 12. FF-NN I and FF-NN II models were evaluated with different predictors.

For FF-NN I, when using the predictors UAW, UUCW, and ECF, the model achieved an *AHP* score of 0.71, with an MDL_{AHP} of -31298.04 and an MDL_{RSS} of 39384.06. However, when using TCF and ECF as predictors, the *AHP* score for FF-NN I dropped to 0.60, with MDL_{AHP} and MDL_{RSS} values of -31921.90 and 40464.73, respectively.

In contrast, FF-NN II with the predictors UAW, UUCW, TCF, and ECF achieved the highest *AHP* score of 1.00, indicating a perfect performance with an MDL_{AHP} of -30093.69 and an MDL_{RSS} of 35846.81. When using only TCF as the predictor, FF-NN II had an *AHP* score of 0.62, and the MDL_{AHP} and MDL_{RSS} were -31764.88 and 40195.76, respectively.

Comparing the models, FF-NN II consistently outperformed FF-NN I across all measures and predictor sets. This suggests that the additional complexity and parameters in FF-NN II provide a better fit for the UCP dataset. The combination of UAW, UUCW, TCF, and ECF yielded the best results for FF-NN II, achieving the highest *AHP* score and the lowest MDL_{RSS}. This combination captures the relevant information more effectively than the other tested sets of predictors. The significant difference in performance measures between the two models and their predictor combinations Table 13 presents the performance of FF-NN models using different predictors combinations on the GLP dataset, evaluated using the *AHP* score, MDL_{AHP}, and MDL_{RSS}. The models compared are FF-NN I and FF-NN II.For FF-NN I, when using the predictors AGE, DBP, SBP, TE, SPO2, HR, SHI, and DN, the model achieved an AHP score of 0.89, with an MDL_{AHP} of -112604.29 and an MDL_{RSS} of 44842.89. However, when using only DN as the predictor, the *AHP* score for FF-NN I dropped to 0.56, with MDL_{AHP} and MDL_{RSS} values of -117876.50 and 50788.08, respectively.

In contrast, FF-NN II with the predictors AGE, DBP, SBP, TE, SPO2, HR, and SHI achieved an AHP score of 0.99, indicating near-perfect performance with an MDL_{AHP} of -111319.67 and an MDL_{RSS} of 42684.87. FF-NN II had an *AHP* score of 0.50 when using only DN as the predictor, and the MDL_{AHP} and MDL_{RSS} were -119184.98 and 50616.86, respectively. Additionally, when using all predictors (AGE, DBP, SBP, TE, SPO2, HR, SHI, and DN), FF-NN II achieved an AHP score of 0.99, with an MDL_{AHP} of -111365.58 and an MDL_{RSS} of 42619.76.

Comparing the models, FF-NN II consistently outperformed FF-NN I across all measures and predictor sets. This suggests that the additional complexity and parameters in FF-NN II provide a better fit for the GLP dataset. The combination of AGE, DBP, SBP, TE, SPO2, HR, and SHI yielded the best results for FF-NN II, achieving the highest *AHP* score and the lowest MDL_{RSS}. This combination captures the relevant health-related information more effectively than the other tested sets of predictors. The significant difference

TABLE 12: AHP, MDL_{AHP}, MDL_{RSS} for FF-NN models with UCP dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
FF-NN I	UAW,UUCW,ECF	0.71	-31298.04	39384.06
FF-NN I	TCF,ECF	0.60	-31921.90	40464.73
FF-NN II	UAW,UUCW,TCF,ECF	1.00	-30093.69	35846.81
FF-NN II	TCF	0.62	-31764.88	40195.76

in performance measures between the two models and their predictor combinations highlights the importance of selecting appropriate predictors. The predictors AGE, DBP, SBP, TE, SPO2, HR, and SHI combined provide a robust model capable of accurately predicting the desired health outcomes in the GLP dataset.

These results demonstrate the efficacy of using a more complex FF-NN model with a comprehensive set of predictors for superior performance in the GLP dataset.

The next dataset GDP resulst are summarised in Table 14. The measures evaluated include the *AHP* Score, MDL*AHP*, and MDL*RSS*. The models were assessed based on different sets of predictors.

For FF-NN I, when using the predictors Interest Rate, Unemployment Rate, Industrial Production, Money Supply, and Personal Income, the model achieved an *AHP* score of -2.86, with an MDL*AHP* of 12.86 and an MDL*RSS* of 1255.61. However, when using only Interest Rate as the predictor, the *AHP* score for FF-NN I remained at -2.86, with MDL*AHP* and MDL*RSS* values of 12.86 and 1255.71, respectively.

In contrast, FF-NN II, with the predictors of Industrial Production and Personal Income, achieved the highest *AHP* score of 1.00, indicating perfect performance with an MDL*AHP* of -294.06 and an MDL*RSS* of 849.81. When using Consumer Sentiment as the predictor, FF-NN II had an *AHP* score of 0.17, and the MDL*AHP* and MDL*RSS* were -415.77 and 1100.61, respectively.

Comparing the models, FF-NN II consistently outperformed FF-NN I across all measures and predictor sets. This suggests that the additional complexity and parameters in FF-NN II provide a better fit for the GDP dataset. The combination of Industrial Production and Personal Income yielded the best results for FF-NN II, achieving the highest *AHP* score and the lowest MDL_{RSS}. This combination captures the relevant economic information more effectively than the other tested sets of predictors. The significant difference in performance measures between the two models and their predictor combinations highlights the importance of selecting appropriate predictors. The Industrial Production and Personal Income predictors provide a robust model capable of accurately predicting the desired economic outcomes in the GDP dataset.

These results demonstrate the efficacy of using a more complex FF-NN model with a comprehensive set of predictors for superior performance in the GDP dataset.

For the last dataset (STOCK) the resulst are in Table 15. Resulst again constisting of scores for the *AHP*, MDL_{AHP}, and MDL_{RSS}

For FF-NN I, when using the predictors High and Low, the model achieved an *AHP* score of 0.83, with an MDL*AHP* of -74826.91 and an MDL*RSS* of -2659.22. However, when using Open, High, and Low as predictors, the *AHP* score for FF-NN I dropped to 0.73, with MDL*AHP* and MDL*RSS* values of -75867.69 and 7134.96, respectively.

In contrast, FF-NN II with the predictor Low achieved the highest *AHP* score of 1.00, indicating perfect performance with an MDL*AHP* of -73376.47 and an MDL*RSS* of -11413.85. When using Date and Open as predictors, FF-NN II had an *AHP* score of 0.83, and the MDL*AHP* and MDL*RSS* were -74826.41 and -2457.73, respectively.

Comparing the models, FF-NN II consistently outperformed FF-NN I across all measures and predictor sets. This suggests that the additional complexity and parameters in FF-NN II provide a better fit for the STOCK dataset. The predictor Low yielded the best results for FF-NN II, achieving the highest *AHP* score and the lowest MDL_{RSS}. This combination captures the relevant stock price information more effectively than the other tested sets of predictors. The significant difference in performance measures between the two models and their predictor combinations highlights the importance of selecting appropriate predictors. The predictor Low provides a robust model that accurately predicts the desired stock price outcomes in the STOCK dataset.

These results demonstrate the efficacy of using a more complex FF-NN model with a comprehensive set of predictors for superior performance in the STOCK dataset.

C. DETAILED DISCUSSION

The comparison between regression models and Multi-Layer Perceptron (FF-NN) models reveals several significant insights across different datasets, as highlighted in Tables 16 and 17. For the GDP dataset, the FF-NN II model achieved an MDL_{AHP} of -415.77, substantially outperforming the Ridge regression model with Interest_Rate+Consumer_Sentiment predictors, which had an MDL_{AHP} of -97.24. Similarly, the FF-NN II model with "Industrial_Production, Personal_Income" showed superior performance with an MDL_{AHP} of -294.06, compared to the Linear Regression model using Interest_Rate+Industrial_Production+Personal_Income, which had an MDL_{AHP} of 10.55.

In the GLP dataset, the FF-NN II model with DN as the sole predictor demonstrated an MDL_{AHP} of -119184.98, significantly better than the Ridge regression model with AGE, which had an MDL_{AHP} of -57652.15. When multiple health indicators were used as predictors, FF-NN II again out-

TABLE 13: AHP, MDL_{AHP}, MDL_{RSS} for FF-NN models with GLP dataset

Model	Predictors	AHP Score	MDL_{AHP}	MDL_{RSS}
FF-NN I	AGE, DBP, SBP, TE, SPO2, HR, SHI, DN	0.89	-112604.29	44842.89
FF-NN I	DN	0.56	-117876.50	50788.08
FF-NN II	AGE, DBP, SBP, TE, SPO2, HR, SHI	0.99	-111319.67	42684.87
FF-NN II	DN	0.50	-119184.98	50616.86
FF-NN II	AGE, DBP, SBP, TE, SPO2, HR, SHI, DN	0.99	-111365.58	42619.76

TABLE 14: AHP, MDL_{AHP}, MDL_{RSS} for FF-NN models with GDP dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
FF-NN I	Interest_Rate,Unemployment_Rate,Industrial_Production,Money_Supply,Personal_Income	-2.86	12.86	1255.61
FF-NN I	Interest_Rate	-2.86	12.86	1255.71
FF-NN II	Industrial_Production, Personal_Income	1.00	-294.06	849.81
FF-NN II	Consumer_Sentiment	0.17	-415.77	1100.61

TABLE 15: AHP, MDL_{AHP}, MDL_{RSS} for FF-NN models with STOCK dataset

Model	Predictors	*AHP* Score	MDL_{AHP}	MDL_{RSS}
FF-NN I	High, Low	0.83	-74826.91	-2659.22
FF-NN I	Open, High, Low	0.73	-75867.69	7134.96
FF-NN II	Low	1.00	-73376.47	-11413.85
FF-NN II	Date, Open	0.83	-74826.41	-2457.73

performed the Ridge regression model, achieving MDL_{AHP} values of -111365.58 compared to -51369.50, respectively. For the STOCK dataset, FF-NN I, using Open, High, Low predictors, achieved an MDL_{AHP} of -75867.69, much better than the Linear Regression's -32535.83. The FF-NN II model using Low alone also performed exceptionally well with an MDL_{AHP} of -73376.47, surpassing the ElasticNet regression model using "Date," which had an MDL_{AHP} of -57585.61.

In the UCP dataset, FF-NN models consistently outperformed regression models. FF-NN I with TCF, ECF achieved an MDL_{AHP} of -31921.90, better than ElasticNet with the same predictors, which had an MDL_{AHP} of -14485.98. The FF-NN II model with UAW, UUCW, TCF, ECF showed an MDL_{AHP} of -30093.69, outperforming ElasticNet and any other regression models tested.

This analysis shows that FF-NN models generally outperform regression models across all datasets in terms of MDL_{AHP}. This suggests that FF-NNs are more capable of capturing complex relationships within the data, which simple regression models might miss. However, while FF-NNs provide significant advantages due to their non-linearity and depth, they also come with higher computational costs and complexity, which can be a disadvantage regarding interpretability and ease of implementation.

Across all datasets and models, MDL_{AHP} consistently provides better performance measure than MDL_{RSS}. For instance, in the GDP dataset, the MDL_{AHP} for FF-NN II with -415.77, while MDL_{RSS} is 1100.61, showing a contrast. This trend is observed across all datasets, underscoring that MDL with AHP is a better method for model selection. It better captures the trade-offs and multi-criteria evaluations inherent in complex model selection, which MDL_{RSS} may oversimplify. However, one must consider that MDL_{AHP} may also involve more subjective judgment in determining weights for different criteria, which can introduce bias.

TABLE 16: Selected FF-NN Models per Dataset and Performance Comparison

Dataset	Model	Predictors	AHP Score	MDL_{AHP}	MDL_{RSS}
GDP	FF-NN II	Consumer_Sentiment	0.17	-415.77	1100.61
GDP	FF-NN II	Industrial_Production, Personal_Income	1.00	-294.06	849.81
GLP	FF-NN II	DN	0.50	-119184.98	50616.86
GLP	FF-NN II	AGE, DBP, SBP, TE, SPO2, HR, SHI, DN	0.99	-111365.58	42619.76
STOCK	FF-NN I	Open, High, Low	0.73	-75867.69	7134.96
STOCK	FF-NN II	Low	1.00	-73376.47	-11413.85
UCP	FF-NN I	TCF,ECF	0.60	-31921.90	40464.73
UCP	FF-NN II	UAW, UUCW, TCF, ECF	1.00	-30093.69	35846.81

TABLE 17: Selected Regression Models per Dataset and Performance Comparison

Dataset	Model	Predictors	AHP Score	MDL_{AHP}	MDL_{RSS}
GDP	LinearRegression	Interest_Rate+Industrial_Production+Personal_Income	0.80	10.55	1066.64
GDP	Ridge	Interest_Rate+Consumer_Sentiment	0.17	-97.24	1409.57
GLP	Ridge	AGE	0.53	-57652.15	111455.88
GLP	Ridge	AGE+DBP+SBP+TE+SPO2+SWE+HR+SHI+DN	0.92	-51369.50	106125.92
STOCK	ElasticNet	Date	0.04	-57585.61	98004.58
STOCK	LinearRegression	Open+High+Low	1.00	-32535.83	22562.16
UCP	ElasticNet	TCF+ECF	0.59	-14485.98	57663.44