applied sciences

MDPI

Article

Machine Learning Ensemble Modelling for Predicting Unemployment Duration

Barbora Gabrikova *, Lucia Svabova * and Katarina Kramarova

Department of Economics, Faculty of Operation and Economics of Transport and Communications, University of Zilina, 010 26 Zilina, Slovakia; katarina.kramarova@fpedas.uniza.sk

* Correspondence: gabrikova1@stud.uniza.sk (B.G.); vrabelova@uniza.sk (L.S.)

Abstract: Predictions of the unemployment duration of the economically active population play a crucial assisting role for policymakers and employment agencies in the well-organised allocation of resources (tied to solving problems of the unemployed, whether on the labour supply or demand side) and providing targeted support to jobseekers in their job search. This study aimed to develop an ensemble model that can serve as a reliable tool for predicting unemployment duration among jobseekers in Slovakia. The ensemble model was developed using real data from the database of jobseekers (those registered as unemployed and actively searching for a job through the Local Labour Office, Social Affairs, and Family) using the stacking method, incorporating predictions from three individual models: CART, CHAID, and discriminant analysis. The final meta-model was created using logistic regression and indicates an overall accuracy of the prediction of unemployment duration of almost 78%. This model demonstrated high accuracy and precision in identifying jobseekers at risk of long-term unemployment exceeding 12 months. The presented model, working with real data of a robust nature, represents an operational tool that can be used to check the functionality of the current labour market policy and to solve the problem of long-term unemployed individuals in Slovakia, as well as in the creation of future government measures aimed at solving the problem of unemployment. The measures from the state are financed from budget funds, and by applying the appropriate model, it is possible to arrive at the rationalization of the financing of these measures, or to specifically determine the means intended to solve the problem of long-term unemployment in Slovakia (this, together with the regional disproportion of unemployment, is considered one of the most prominent problems in the labour market in Slovakia). The model also has the potential to be adapted in other economies, taking into account country-specific conditions and variables, which is possible due to the data-mining approach used.

Keywords: unemployment; ensemble modelling; CRISP-DM; data-mining; unemployment duration prediction

check for updates

Citation: Gabrikova, B.; Svabova, L.; Kramarova, K. Machine Learning Ensemble Modelling for Predicting Unemployment Duration. *Appl. Sci.* **2023**, *13*, 10146. https://doi.org/10.3390/app131810146

Academic Editor: Donghai Guan

Received: 15 August 2023
Revised: 3 September 2023
Accepted: 6 September 2023
Published: 8 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Unemployment is a pervasive social and economic problem that affects individuals, families, and communities across the globe. It can result in reduced economic growth, increased social inequality, and negative mental health and well-being impacts. Its negative effects can be long-lasting and far-reaching [1–3], and Slovakia is not an exception.

Unemployment in Slovakia is a long-discussed issue and, from the point of view of the government, a priority issue to solve. In Slovakia, the unemployment rate has been relatively persistently high; however, the existing problem of unemployment is not a uniform matter. Mainly long-term unemployment and structural unemployment, according to the qualifications of the workforce and to regional aspects (so-called "hungry valleys"), pose a particular challenge to solve.

Addressing unemployment is a complex and multifaceted challenge requiring various strategies and interventions. Some potential solutions include improving access to education and training programs to equip jobseekers with the skills needed to succeed in the job market. Additionally, promoting entrepreneurship and small business development can create more job opportunities. Furthermore, implementing policies that support economic growth and job creation is crucial [4–6]. Globally, effective strategies to address unemployment require collaboration between governments, employers, and civil society organisations. They must develop co-ordinated and comprehensive approaches that address the underlying causes of unemployment and support jobseekers in finding and maintaining employment [7,8].

Despite various policy efforts to address this issue, there remains a need for more effective solutions to reduce the negative impact of unemployment on individuals and society. To address this issue, researchers have explored various approaches, including the use of machine learning models, to predict unemployment duration for jobseekers [9,10].

Predictions of the unemployment duration of jobseekers can be a useful tool for policymakers, employers, and jobseekers themselves. Machine learning models can identify relevant factors indicating unemployment duration, enabling the development of targeted interventions supporting jobseekers to find an appropriate job more quickly. These interventions may include, e.g., job training programs, employment incentives, and job matching services that connect jobseekers with suitable job opportunities. Policymakers can implement targeted job training programs, employers can inform recruitment strategies, and, finally, jobseekers can make informed decisions about their job search, financial planning, and personal well-being [6,11].

Overall, predictions of unemployment duration may help mitigate the negative effects of unemployment on individuals and society, as well the economic performance of the national economy (since unemployment affects private consumption with a possible impact on the real performance of the economy). It also supports the development of more effective policies and interventions to address this persistent social and economic challenge.

We view machine learning modelling as a promising approach for predicting unemployment duration and developing targeted interventions supporting jobseekers. The field of machine learning is rapidly advancing, aiming to develop algorithms capable of learning from data and making predictions or smarter decisions based on that learning. Machine learning is related to artificial intelligence. It deals with computer systems and algorithms that can solve specific tasks consisting of complex processes, based on learning from the provided data, without pre-programmed rules. Machine learning approaches are now used in various areas and applications, such as image and speech recognition, natural language processing, as part of internet offers and sales (recommender systems), within banking and financial services (e.g., the detection of unusual financial transactions), within accounting and systems to uncover tax fraud, within medical and pharmaceutical processes, within transport and logistics (e.g., autonomous vehicles and the automation of logistics processes), the optimization of energy infrastructure, the optimalization of management, the optimization of various areas of business management (e.g., predictions of financial health, the optimization of supply processes, storage, the optimization of targeting of marketing tools, and the optimization of investment decisions), etc. [12,13].

This research article contributes to the field of unemployment analysis, management, and solving, as well as the analysis, control, and preparation of intervention strategies by exploring the use of ensemble machine learning models in the modelling of the duration of the unemployment of jobseekers (officially registered jobseekers) in Slovakia. This research aims to obtain predictions that are as accurate as possible of how long a jobseeker will likely be unemployed based on his/her own specific characteristics and unemployment history. By identifying the key factors that influence the duration of the status “the unemployed”, this study provides insights that can be helpful for preparing more effective policies and interventions to support jobseekers, and consequently help to reduce the social and economic burdens of unemployment in Slovakia.

Literature Review

In the scientific literature, various methods have been employed to predict the unemployment rate and its development, or possibly other quantitative characteristics of the labour market. Generally, predicting the phenomena occurring in this market helps the entities involved there to verify and initiate a development lane aimed at meeting the expected state in the future.

In the context of the research article presented by us, we can point out, e.g., the study by [9], where the author explored time series and machine learning models such as the extended version of the FARIMA model, which incorporates the conditional heteroskedasticity of errors (FARIMA/GARCH), as well as artificial neural networks, support vector regression, and multivariate adaptive regression splines for the prediction of unemployment in the selected European countries and for various forecasting periods. When comparing the results obtained from the individual models, the author found out that selecting a specific approach to unemployment rate modelling should take into consideration not only the geographical location but also the time horizon of the forecast. Another study presented by [14] focused on forecasting the unemployment rate using machine learning methods, with the neural network providing the best results of the predictions. The approach based on the neural networks was also employed by [15]. The authors used recurrent neural networks to predict two macroeconomic variables—unemployment rate and inflation in the USA. The results indicate that longer patterns are more suitable for predicting unemployment, while smaller patterns are more suitable for predicting inflation. The results also indicate that adding more layers does not necessarily guarantee more accurate predictions. In [16], the authors forecasted unemployment in China by employing neural networks and the ARIMA time series model. They combined the results from both methods to enhance the prediction accuracy, creating an ensemble combined model. In [17], the author employed the Cox proportional hazards regression model to identify the factors influencing the duration of unemployment of jobseekers in Ukraine, working with the data on unemployed individuals from 1998 to 2002. The results of the study indicate the same problems of unemployment as in our country, e.g., the existence of disadvantaged groups of unemployed people with respect to the probability of re-employment, such as a low-skilled workforce, older individuals, and individuals living in rural areas and small towns (i.e., geographical disproportions in unemployment). Similarly, in [18], the authors examined the factors impacting the duration of unemployment of jobseekers in Turkey using survival analysis. This approach was chosen mainly because it made it possible to identify social, demographic, and economic factors that influence the duration of unemployment of jobseekers. Considering the severe psychosocial and economic consequences of long-term unemployment after university graduation, the authors of [19] employed a logistic regression model to examine the factors influencing the duration of this type of unemployment in Rwanda. The key factors identified include the graduate's age, job-seeking methods, and acquired skills. Likewise, the study of [19] utilised a logistic regression model to uncover the determinants of youth unemployment in Ethiopia. Furthermore, a study [20] deals with identifying factors associated with low graduate employment in Malaysia. In [21], the authors employed panel data analysis to examine the key determinants of youth unemployment across 31 OECD countries. The study selected GDP growth rate, inflation rate (measured using the consumer price index), gross domestic savings (expressed as a percentage of GDP), and labour productivity growth as independent variables. Among these variables, only labour productivity positively affects youth unemployment, while the remaining independent variables have a negative effect. Notably, the variable that most significantly influences youth unemployment is the GDP growth rate.

A panel data analysis, namely dynamic, was also applied in [22], where the impact of selected macroeconomic factors on the unemployment rate in the EU member states was analysed and assessed. Also, the study by [23] focused on the analysis of the EU member states' unemployment, however with more details on the examination of the selected macroeconomic and structural factors influencing, namely, youth unemployment from 2008 to 2018. Different machine learning techniques such as random forests, support vector machines, elastic-net logistic regression model, and classification and regression trees (CARTs)—the method also applied by us—were used to predict, e.g., the Eurozone unemployment rate by [24]. Similar machine learning techniques, including discriminant analysis, the other method applied in our study, were presented in [25] to determine the primary causes of unemployment in the USA using data from 1976 to 1986. By applying these techniques, among other things, the authors acquired important information on the geographical structure of US unemployment by region. The status and needs of the US labour market was analysed in the study by [26], too. The study assesses the selected work incentive and suggests a possible methodology for its impact valuation (on employment outcomes) employing CHAID analysis, i.e., the same method applied by us. By applying the same procedure in the same labour market, the methods of discrimination against the employment of jobseekers with disabilities were investigated in the study by [27]. This type of jobseeker represents a part of the labour force that constitutes a specific part of the total unemployed population in each economy; it is also characteristic for Slovakia. In the study by [28], another machine learning technique to predict the development of the unemployment rate was used (average window forecasts), as well as standard ARMA-based time series approaches. The overall aim of this study was to find the most useful framework for time series predicting via machine learning.

In Slovakia, ref. [29] investigates the impact of various factors, including GDP per capita, overall unemployment rate, apartment price per square meter, and others, on the unemployment rate of high school graduates. Two logistic regression models were developed to examine the influence of these factors on the unemployment of this kind of workforce. The results revealed a statistically significant relationship between the unemployment of high school graduates and the overall unemployment rate in the region, the GDP per capita in the region, the quality of secondary education, and the cost of living in the region immediately after graduation. Furthermore, ref. [30] examines the main factors affecting structural unemployment in Slovakia between 2014 and 2019. Using a panel regression model and its modification, the study analyses the influence of factors on the evolution of the structural unemployment rate. Based on the findings, a decomposition of the unemployment rate in the V4 countries was subsequently conducted, identifying important cyclical and structural factors for each country. The created model effectively captures a significant portion of the unemployed during the observed period. In addition, a study [31] focuses on forecasting the time series of unemployment development in Slovakia and utilises ARIMA and GARCH models.

Currently, some studies leverage Internet search data to predict unemployment. In [32], the author utilises MIDAS regression models to forecast unemployment and evaluate the utility of Google search data as a predictor in the USA. Similarly, in [33], the authors investigated the viability of using Google search data to predict unemployment in Spain. In [34], the authors employed the PRISM semiparametric method to examine the predictive value of Internet search data for unemployment. This method demonstrates superior predictive capability not only compared to approaches used during the financial crisis era but also surpasses predictions made during the COVID-19 pandemic.

2. Methodology and Data

Ensemble modelling, utilising machine learning algorithms, has emerged as a technique to enhance the accuracy and robustness of predictive models. This approach involves combining multiple models, often trained on different subsets of data or using different algorithms, to produce more accurate and reliable predictions. One of the main advantages of this approach is its ability to reduce the risk of model overfitting [35,36].

Ensemble modelling can be classified into two main types: bagging and boosting. *Boosting* entails training a sequence of models, where each subsequent model is trained to correct the errors of the previous model. On the other hand, *bagging* involves training multiple models independently on different subsets of the training data [37,38]. The individual predictions from these models are then combined through aggregation. The fundamental concept behind aggregation is to create multiple models that each provide a prediction on the same set of inputs. These individual predictions are then combined to generate a final prediction, resulting in a more accurate and robust prediction compared to that of any single model [35].

Various methods are available for combining models, and the choice of technique depends on the specific problem and the types of models being used. The most commonly used techniques for combining models include the following [35,39,40]:

- *Stacking*: This method involves training multiple models and then using a meta-model to combine their predictions. The meta-model takes the predictions from each base model as inputs and learns to weigh them appropriately to produce the final prediction;
- *Competition*: In this method, individual predictions are considered, and the prediction with the highest confidence is selected for each case in the data;
- *Voting*: This method combines the individual models by taking the average of their predictions;
- *Weighted Voting*: Similar to the voting method, this approach also involves taking the average of the predictions from each model. However, the predictions are weighted based on a specific metric, such as confidence.

Overall, combining models is a powerful technique for improving the accuracy, precision, and robustness of predictive models. By leveraging the strengths of multiple models, ensemble methods can often outperform any single model and provide more reliable predictions [41].

In this article, we aimed to create an ensemble model for predicting the unemployment duration of jobseekers in Slovakia. To achieve this, we employed the stacking method to combine the base models. The unemployment duration was quantified on a binned scale, commonly used in practice for unemployment evaluation (e.g., [42,43]). We categorised the unemployment duration into the following categories: up to 3 months, 3–6 months, 6–12 months, and more than 12 months. Consequently, this prediction task falls under supervised machine learning, where the algorithm learns from labelled data with the known target variable.

As part of the supervised learning machine learning techniques, we utilised *discriminant analysis* and two decision tree types: the *Classification and Regression Tree* (CART) and *Chi-squared Automatic Interaction Detection* (CHAID). Each of these techniques possesses distinct advantages and weaknesses. Combining their predictions can produce more precise results than any individual model's prediction. The final model was created through *stacking*, where the individual predictions of the models mentioned above were used as inputs for the final meta-model. As the final meta-model, we employed *logistic regression*, with predictions based on the predictions of the individual models.

We chose this method because it not only predicts the classification of the case into one of the categories of unemployment duration, but also predicts the probability for each category. We consider the prediction of probabilities an advantage because it allows for the customisation of classification based on the predicted probabilities and the type of the solved issue. That means that, instead of relying on equally divided probabilities, the jobseeker could be classified into one of the categories of predicted unemployment duration using the expertly set dividing criteria.

All these machine learning methods were selected because of their interpretability of results [44]. Other methods, such as neural networks, nearest neighbours, or random forests, can provide highly accurate predictions, but they are often regarded as "black boxes" due to the lack of interpretability of their models. In this study, we aimed to have readable and interpretable results, not only for the prediction of the individual unemployment duration but also for identifying its main determinants. To achieve this, we calculated the importance of the predictors in each prediction model.

In the text bellow, we introduce a concise overview of the main principles, advantages, and disadvantages of the applied machine-learning methods.

2.1. CART Decision Tree

Classification and Regression Trees (CARTs) are a widely used machine learning algorithm for classification and regression analysis, first introduced by Breiman, Friedman, Olshen, and Stone in 1984 [45]. The CART is a non-parametric decision tree method that uses a recursive partitioning strategy to create a tree-based model.

The algorithm begins with the entire dataset and then selects the variable and its threshold that best separates the data into two subsets based on the impurity of the resulting subsets. The impurity can be measured using different metrics; the mostly used impurity measures are the Gini index or entropy. The whole tree structure is then created by repeatedly splitting the data into two smaller subsets based on the values of the input variables. The splitting process is repeated until a stopping criterion is met, such as a minimum number of observations per leaf or a maximum tree depth [46,47]. In this study, we used a maximum tree depth of 5; minimum records in the parent branch of 2%; minimum records in the child branch of 1%; and a minimum change in impurity of 0.0001 as the stopping criteria. The impurity is measured using the Gini index, which measures the probability of misclassifying a randomly chosen data case from a given node. It is calculated as follows:

$$\text{Gini Index} = 1 - \sum {p_i}^2$$

where p_i is the proportion of samples in class i at the node. The Gini index reaches its minimum when all samples at the tree node belong to the same class of the target variable.

Entropy measures the amount of information or uncertainty in a given data set. It is calculated using

$$\text{Entropy} = -\sum p_i \cdot \log p_i$$

where p_i is the proportion of samples in class i at the node. Entropy reaches its minimum when all samples at the node belong to the same class.

The Gini index and entropy are the most commonly used impurity measures in the CART algorithm. Some studies have shown that the Gini index tends to perform slightly better in practice for classification tasks, while entropy may be more sensitive to changes in the data distribution [37,38].

The CART method can handle both categorical and continuous input variables, which is considered one of its main advantages. CART modelling results are unaffected by either the collinearity between the input variables or the occurrence of outliers in the dataset. It can handle the missing data by setting the surrogate input variables for each variable. As the main disadvantage of the method, we mention the problem of overfitting, where the tree is too complex and fits the noise in the training data instead of the general patterns [45]. To solve this problem, the CART could be pruned after its building to reduce overfitting and improve generalisation performance. Pruning means removing such branches of the tree that do not significantly improve the overall accuracy of the tree [45,48]. To avoid this problem of overfitting, the CART in this study was pruned.

2.2. CHAID Decision Tree

CHAID (Chi-squared Automatic Interaction Detection) is a decision tree algorithm suitable for supervised classification or regression learning with categorical target variables. It was introduced in the study by [49]. In this study, we used it for classification tasks. Compared with other decision tree algorithms, CHAID has some unique features; for example, it allows splitting the node into more than two subsets, leading to more accurate and interpretable models [50].

The algorithm begins with the entire dataset and then selects the predictor variable and category that best separates the data into two subsets based on the chi-squared statistic. The data are split into two or more subsets based on the values of that variable. The significance of each split is evaluated using the chi-square statistic, which measures the independence between the response variable and each predictor variable. In more detail, the chi-squared statistic is used to evaluate the statistical significance of the differences in the target variable across the categories of the predictor variable. If the chi-squared statistic is significant at a given significance level, the split is considered valid, and the algorithm proceeds to split the data further. If it is insignificant, the algorithm stops and considers the current node a leaf node [51].

The process is repeated until a stopping criterion is met, such as a minimum number of observations per leaf or a maximum tree depth. In this study, we used the following stopping criteria: a maximum tree depth of 5; minimum records in the parent branch at 2% and in the child branch at 1% of cases in the training sample; a minimum change in the expected call frequencies for the chi-square test of 0.001; maximum iterations for convergence of 100. Similarly, as in the CART method, the tree can be pruned to reduce overfitting and improve generalisation performance, as the possibility of overfitting belongs to its main disadvantages.

2.3. Discriminant Analysis

Discriminant analysis is a statistical technique used to identify differences between two or more groups based on a set of predictor variables. The technique aims to find a linear combination of the predictor variables that maximally discriminate between the k groups of the outcome variable. This linear combination is known as the discriminant function $f(x)$. Its formula is as follows [52,53]:

$$f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

where p is the number of predictor variables, $x = \left(x_1,\ x_2, \ldots,\ x_p\right)$ is an observation of the p predictor variables, and $\alpha_0, \alpha_1, \alpha_2,\ \ldots, \alpha_p$ are parameters to be determined; their estimations will be $a_0, \alpha_1, a_2, \ldots, a_p$.

The goal is to find the estimates $(a_0, \alpha_1, a_2, \ldots, a_p)$ of the coefficients $\alpha_0, \alpha_1, \alpha_2,\ \ldots, \alpha_p$, such that the discriminant function $f(x)$ separates the groups of the target variable as well as possible. One common approach is maximising the discriminant function of between-group variance relative to the within-group variance. Specifically, we seek to maximise the ratio of the between-group variance to the sum of the within-group variances, known as Fisher's discriminant ratio.

For linear discriminant analysis, the solution can be expressed in terms of the inverse of the pooled covariance matrix, which is a weighted average of the group covariance matrices. We assume that the covariance matrices are equal across groups for the linear discriminant analysis but not necessarily for the quadratic discriminant analysis. For the quadratic discriminant analysis, we need to estimate separate covariance matrices for each group, and the solution is given using a quadratic function of the predictor variables [52–54].

When the target variable has K possible values, this involves fitting K into separate discriminant functions, one for each class, where each function is trained to distinguish that class from all other classes of the outcome. Specifically, for class k, we define a binary response variable of y_k, where $y_k = 1$ if the observation belongs to class k, and $y_k = 0$ otherwise. We then fit a discriminant function of $f_k(x)$ for each k, using the same predictor variables as before. Finally, to classify a new case, we compute the values of all K discriminant functions and assign the case to the class with the highest value of $f_k(x)$.

One of the key advantages of discriminant analysis with multiple target variables is that it allows for identifying the most important predictors of group membership. Additionally, this method can identify which variables are the most influential in discriminating between groups, providing insight into the underlying mechanisms driving group differences [55].

2.4. Logistic Regression

Logistic regression is a statistical technique used to model the relationship between a binary or multinomial target variable and predictor variables. The target variable is modelled as a function of predictor variables using the logistic function, which transforms the linear predictor into a probability between 0 and 1 [56].

For the case of a binary target variable, the principle of logistic regression is the following: let Y be a binary response variable that takes on the values of 1 with probability p and 0 with probability $1 - p$. Let $X_1, X_2, \ldots, X_m$ be m predictor variables. The logistic regression model expresses the log odds of the target variable Y being equal to 1 as a linear function of the predictor variables:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ are the parameters to be estimated, and their estimations will be $b_0, b_1, b_2, \ldots, b_m$. Then, to transform the linear predictor into a probability between 0 and 1, the logistic function is used:

$$p = 1/(1 + exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_m X_m))$$

To estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$, the maximum likelihood method is used. The likelihood function is given using

$$L(\beta_0, \beta_1, \beta_2, \ldots, \beta_m) = \prod {p_i}^{y_i} \cdot (1 - p_i)^{1-y_i}$$

where the product is taken over all i observations, y_i is the observed value of the outcome variable (0 or 1) for the i-th observation, and p_i is the predicted probability of the outcome variable being equal to 1 for the i-th observation.

The maximum likelihood estimates of the parameters can be found using iterative methods such as the Newton–Raphson method or the Fisher scoring algorithm. The parameter estimates are obtained by maximising the log-likelihood function [57,58]:

$$\ln L(\beta_0, \beta_1, \beta_2, \ldots, \beta_m) = \sum (y_i \ln p_i + (1 - y_i)\ln(1 - p_i))$$

The methodology could be extended intuitively for the multinomial target variable, as is the categorised duration of unemployment in this study. The multinomial logistic regression model assumes that the log odds of the outcome variable Y taking on category k is a linear function of the predictor variables:

$$\log\frac{p(Y = k)}{p(Y = K)} = \beta_{0k} + \beta_{1k} X_1 + \beta_{2k} X_2 + \cdots + \beta_{mk} X_m$$

where $p(Y = k)$ is the probability that the outcome variable Y will have the value k; $\beta_{0k}, \beta_{1k}, \beta_{2k}, \ldots, \beta_{mk}$ are the parameters to be estimated for category k, their estimations will be $b_{0k}, b_{1k}, b_{2k}, \ldots, b_{mk}$, and K is the total number of categories.

The multinomial logistic function is then used to transform the linear predictor into a probability distribution across the categories:

$$p(Y = k) = \exp(\beta_{0k} + \beta_{1k} X_1 + \beta_{2k} X_2 + \cdots + \beta_{mk} X_m) \\ / \sum (\exp(\beta_{0j} + \beta_{1j} X_1 + \beta_{2j} X_2 + \cdots + \beta_{mj} X_m)$$

where the sum is taken over all possible categories of the target variable.

The maximum likelihood method estimates the parameters that maximise the likelihood of observing the data given in the model. The likelihood function is given using

$$L(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \prod (p(Y_i = k))^{y_i}$$

where the product is taken over all i observations, y_i is the observed value of the outcome variable for the i-th observation, and $p(Y_i = k)$ is the predicted probability of the outcome variable taking on category k for the i-th observation. The parameter estimates are obtained by maximising the log-likelihood function:

$$ln(L(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)) = \sum(y_i \ln p(Y_i = k))$$

In this manner, we obtain the estimated probability for each case in the data to belong to the k-th value of the outcome variable. The case is then predicted to have the k-th value of the outcome variable when this probability is the highest. The maximum likelihood estimates of the parameters can be obtained using iterative methods such as the Newton–Raphson method or the Fisher scoring algorithm [57,58].

2.5. Data

In this study, we analyse the unemployed jobseekers (i.e., officially registered jobseekers) in Slovakia and predict the duration of their unemployment, which we sort into categories according to the duration of the status "unemployed", as we mentioned above.

To create the prediction models, we worked with the robust data coming from the database of jobseekers administered by the governmental public employment agency the Central Office of Labour, Social Affairs and Family of the Slovak Republic (COLSAF). Registration in this database is mandatory for all unemployed jobseekers who are provided with related benefits such as unemployment allowance, participation in intervention programs, or state-paid health insurance, all in accordance with Act No. 5/2004 Coll. on employment services, as amended [59]. The database of jobseekers used in this study covers the period from April 2010 to June 2019. This restriction is due to the complexity of the data available from the COLSAF.

The dataset contains 1,048,551 registrations from 718,032 jobseekers registered in the database once or multiple times during the observed years. We worked with specific characteristics of jobseekers, i.e., qualitative and quantitative information requested from each jobseeker according to the law in force, namely [59], and that are usually detected during the process of registration or during the recording of jobseekers. All mentioned information is included in the register only on the basis of notification and the persons concerned who provide it. Taking into account these characteristics, which are detected when registering a jobseeker or already during his/her existing registration, the research question we set is whether the relevant variables (characteristics of an individual as a jobseeker) are suitable predictors of how long the respective individual will be "about unemployed".

The variables selected by us as being suitable for our study, namely qualitative characteristics used as statistical variables with their short but exact description and with the distribution of their values in the dataset, are presented in Table 1, while the quantitative variables and their characteristics are provided in Table 2.

Table 1. Qualitative input variables used in the study.

Variable	Description	Values	Proportion
gender	gender of jobseeker	male	52.26%
		female	47.74%
nationality	nationality of jobseeker	Slovak	89.31%
		Hungarian	9.35%
		unknown or other	0.80%
		Czech	0.37%
		Roma	0.16%

Table 1. *Cont.*

Variable	Description	Values	Proportion
marital status	marital status of jobseeker	single married divorced widow unknown	51.51% 37.21% 9.36% 1.50% 0.42%
permanent residence	permanent residence of the jobseeker (part of Slovakia)	Eastern Slovakia Western Slovakia Central Slovakia Bratislava region	33.79% 32.13% 26.72% 7.36%
education	highest achieved education of the jobseeker	non-finished primary primary lower secondary vocational secondary vocational complete secondary general secondary higher vocational university 1st university 2nd university 3rd NA	28.32% 30.03% 13.04% 5.16% 0.72% 0.17% 8.86% 0.57% 2.55% 9.25% 1.31%
disadvantage: school leaver	disadvantaged jobseeker	no yes	84.53% 15.47%
disadvantage: over 50 years	disadvantaged jobseeker	no yes	84.08% 15.92%
disadvantage: long-term unemployed	disadvantaged jobseeker	no yes	68.18% 31.82%
disadvantage: health	disadvantaged jobseeker	no yes	97.13% 2.87%
disadvantage: no paid job	disadvantaged jobseeker	no yes	55.60% 44.40%
disadvantage: low education	disadvantaged jobseeker	no yes	86.19% 13.81%
disadvantage: organisational reasons	disadvantaged jobseeker	no yes	99.58% 0.42%
disadvantage: others	disadvantaged jobseeker	no Yes	99.70% 0.30%
children	number of children of the jobseeker	0 1 2 3 4 or more	87.78% 6.57% 3.85% 1.16% 0.64%
reason of exclusion	reason of exclusion from the database of jobseekers	employment relevant reason of exclusion non-co-operation	54.25% 34.78% 10.97%
intervention	intervened jobseeker	No Yes	65.95% 34.05%
previous registrations	number of all previous registrations	0 1 2 3 4 5 or more	64.94% 22.51% 7.02% 2.47% 1.11% 1.95%

Table 1. *Cont.*

Variable	Description	Values	Proportion
previous non-interventions	number of previous registrations without intervention	0 1 2 3 or more	87.54% 10.25% 1.79% 0.42%
previous interventions	number of previous registrations with intervention	0 1 2 3 4 or more	87.81% 6.03% 2.29% 1.21% 2.67%

Table 2. Quantitative input variables used in the study.

Variable	Description	Mean	Min	Max	Range	Std. Deviation	Median	Mode
age	age of jobseeker at the beginning of registration	34.9	15.0	78.0	63.0	12.5	32.0	24.0
cumulative previous registrations	cumulative number of days of previous registrations before the period under review	1155.1	0.0	33,722.0	33,722.0	1343.8	699.0	0.0
work before registration	number of days from the last occupation to the current registration	414.7	0.0	15,805.0	15,805.0	1068.4	1.0	0.0
average unemployment	average number of days spent in unemployment per year from 15 years of age	45.07	0.0	364.94	364.94	54.2	26.1	0.0
duration of registration	duration of the current registration in days	300.9	0.0	1365.0	1365.0	303.8	182.0	91.0

Table 2 contains basic descriptive characteristics, such as the *mean* (average value of the variable), *median* (middle value among the values arranged in ascending order), *mode* (value with the most frequent occurrence), *minimum* and *maximum* value, *range* (difference between the maximum and minimum values), and *standard deviation* (basic characteristic of the variability and square root of the variance, which is an average square difference from the mean).

The last row of Table 2 presents the characteristics of the target variable before its categorisation into the mentioned four categories according to the duration of the status "unemployed". We would like to emphasise that each case in the dataset represents one individual registration in the database. It is important to note that some jobseekers have multiple registrations during the period under review (see variable "previous registrations"; e.g., almost 65% of jobseekers were registered just once, the rest more than once). Therefore, we included variables such as the number of previous registrations, the number of previous non-interventions, the number of previous interventions, the cumulative previous registrations in days, and the average unemployment in days in the data. These variables capture the number or duration of all previous registrations or interventions in which jobseekers participated, as they are expected to impact their unemployment.

All the variables mentioned above (both qualitative and quantitative) will serve as inputs for the prediction models of unemployment duration, i.e., the target variable in this study is the *duration of registration* in the database of jobseekers. This duration is calculated as the difference between the start date and the end date of registration. For jobseekers whose registration had not ended at the time of data extraction, we considered the last day of the period under review as the end date to establish their duration of unemployment as well. The difference in dates was calculated in months and then categorised into the already-mentioned categories: up to 3 months, 3–6 months, 6–12 months, and more than 12 months.

The distribution of the categorised target variable values among jobseekers in Slovakia during the period under review is presented in Table 3. To enhance the model's precision and confidence, we applied boosting to balance the subsamples based on the target variable. This involved balancing all subsamples to have approximately the same frequency as the most numerous group of the target variable while boosting the other groups by assigning appropriate weights to the cases. The weights used for balancing are also provided in Table 3.

Table 3. Distribution of target variable "duration of registration", i.e., the status "unemployed".

Duration of Registration	Proportion of the Sample	Multiplication Factor for Balancing via Boosting
up to 3 months	23.85%	1.2036
3–6 months	21.34%	1.3453
6–12 months	26.10%	1.0998
more than 12 months	28.71%	1.0000

From this partial analysis, it specifically follows that the long-term unemployed people represent the largest group of job seekers (note: in the case of Slovakia, the term long-term unemployed defines a person who is defined as an economically active person, but has not worked for more than 1 year, i.e., was registered in the database of job seekers during this period). On the other hand, it can be concluded that more than 70% of jobseekers were registered as unemployed for less than 1 year, i.e., the period under review was characterized by the persistence of short- and medium-term unemployment.

2.6. Evaluation of Results

To evaluate the results obtained, we employed various evaluation statistics. To ensure the relevance of evaluating the quality of the created ensemble model, we divided the dataset into a training part, on which the machine learning models were created, and a testing part, which served for the independent evaluation of the predictions. Therefore, predictions of unemployment duration were generated for all units in the dataset, including those in the testing set. Predictions of registrations in the testing set were made using an ensemble model that had not been trained, which allowed us an independent assessment of the quality of the model by comparing these predictions to actual outcomes.

In this study, we only present the evaluation results for the testing set. Given enough cases in the dataset, the training and testing parts of the sample were allocated in a random 50:50 ratio.

We utilised the statistics mentioned in the methodology section to evaluate the individual results. The entire ensemble model was initially evaluated using a confusion (or coincidence) matrix and its associated statistics, i.e., let TP represent the number of true positive cases; FP the number of false positive cases; TN the number of true negative cases; and FN the number of false negative cases, where the "hit" is the importance value of the target variable. The overall model evaluation was based on its *accuracy*, which measures the proportion of correctly classified cases from the entire testing sample [60]:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}. \tag{1}$$

For a more detailed view, we also report the following [60]:

- *sensitivity*—the row % of true positive predictions, calculated as follows:

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

- *specificity*—the row % of true negative predictions, calculated as follows:

$$Spec = \frac{TN}{FP + TN} \tag{3}$$

- *precision*—the column % of true positive predictions, calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{4}$$

- *F-measure*—the harmonic mean of the accuracy and sensitivity, calculated as follows:

$$F = \frac{2TP}{2TP + FP + FN}. \tag{5}$$

Since the target variable consists of four possible values instead of being binary, the calculation of these evaluation measures must be adjusted. The general coincidence matrix cannot be used directly and thereby we employed the *macro-average* and *micro-average* methods. These approaches involve converting the general coincidence matrix into a 4-field matrix. We performed this conversion for each category of the target variable, comparing the prediction for each category against the summed predictions for all other categories. The evaluation measures (1)–(5) were then calculated based on the resulting 4-field tables, with either the average (macro-average) or by summing partial 4-field tables and calculating the evaluation measures from the resulting combined table (micro-average).

To assess the importance of the individual input variables in the individual models, we calculated the predictor importance. This analysis aims to highlight the characteristics of an individual that have the greatest impact on the duration of jobseekers' unemployment.

All calculations were performed using IBM SPSS Modeler 18.3, a data-mining software. We approached this problem as a data-mining task, working with large datasets (big data) that represent the entire population of registered jobseekers during the monitored period rather than a random sample from the population. It is important to note that statistical tests of hypotheses about the models or the significance of variables are often not used in data-mining machine-learning models. Therefore, we did not employ stepwise methods based on the statistical testing of the significance of the variables to create the models (discriminant analysis and logistic regression). Instead, the focus was on the predictions generated by the ensemble model. However, we used the significance of the variables to interpret the importance of the individual factors for the duration of unemployment of jobseekers. The interpretability of the results was a key consideration in selecting the machine-learning techniques for modelling, as mentioned earlier.

The outputs of the individual classification trees are presented as dendrograms expressed as if–then rules. However, due to the extensiveness of the results, we have included them in the appendix solely in the form of if–then rules. This decision was made because editing the resulting tree dendrograms would render them unreadable. The discriminant analysis model is provided in the appendix as a set of discriminant Fisher equations, which are used to calculate the discriminant score. Additionally, the table containing the results of the logistic regression meta-model is included directly in Section 3.

3. Results

In this section, we provide the primary findings of our analyses and describe the models developed for predicting unemployment duration. Initially, we concentrate on the individual models, which were subsequently employed to construct the final logistic regression meta-model.

3.1. CART Model

The CART model was constructed using a boosting technique to enhance model accuracy. By employing this method, the algorithm generated 10 component models, which were then utilised to train the boosted model. The individual models achieved accuracies (Formula (1)) ranging from 25 to 53.2%. The ensemble CART model achieved an accuracy of 53.1%. For comparison, the naïve model, which randomly predicts one of the four categories of unemployment duration, has an accuracy of such random prediction of 25%. In this sense, the accuracy of the boosted ensemble CART model of 53.1% indicates that over half of the registrations of jobseekers were correctly classified into one of four categories of unemployment duration. We note that this accuracy pertains to the training set. To calculate the evaluation statistics (1)–(5) for the testing set, we computed the micro- and macro-average as explained in the methodology section.

For a more comprehensive understanding of the jobseekers' correctly and incorrectly classified registrations in the testing set, we provide the coincidence matrix presented in Table 4.

Table 4. Coincidence matrix of the boosted CART model.

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Sum	Correct [%]
up to 3 months	86,484	10,519	25,631	2391	125,025	0.692
3–6 months	54,710	25,734	29,698	1890	112,032	0.230
6–12 months	64,445	8266	60,991	2974	136,676	0.446
more than 12 months	11,136	2691	22,551	114,202	150,580	0.758
sum	216,775	47,210	138,871	121,457	524,313	0.548

As mentioned above, to evaluate the sensitivity, specificity, precision, and F-measure of the model, we recalculated the confusion matrix into a 4-field matrix for each category of the target variable. These individual tables enable us to assess the performance of the model for each unemployment duration category. We present the cumulative 4-field confusion matrix in Table 5 and all values of the evaluation measures for the boosted CART are presented in Table 6. In the case of the evaluation measures, we do not only present the information for individual period categories of the length of the status "unemployed" (first four columns of Table 6), but also their macro- and micro-averages across the entire CART model (last two columns of Table 6).

Table 5. Cumulative confusion matrix for boosted CART ensemble model.

Unemployment Duration	Sum Correct	Sum Incorrect	Sum	Correct [%]
sum correct	287,411	236,902	524,313	**54.82**
sum incorrect	236,902	1,336,037	1,572,939	**84.94**
sum	524,313	1,572,939	2,097,252	**77.41**
correct [%]	**54.82**	15.06		

Table 6. Evaluation measures for boosted CART model.

Evaluation Measure	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Macro-Average	Micro-Average
sensitivity	69.17	22.97	44.62	75.84	53.15	54.82
specificity	67.37	94.79	79.91	98.06	85.03	84.94
precision	39.90	54.51	43.92	94.03	58.09	54.82
F-measure	50.61	32.32	44.27	83.96	52.79	54.82
accuracy	39.90	79.44	70.71	91.68	77.41	77.41

All of these measures were calculated for the testing set to determine the true quality of the model, unaffected by overfitting. Following the results, the boosted CART model demonstrated the best performance in determining the long-term unemployed category of jobseekers (those unemployed for more than 12 months), with a sensitivity of 75.8%. This implies that the model correctly predicted 3/4 of individuals who were unemployed for longer than 12 months to belong to the long-term unemployed category based on their input values. Furthermore, this category had the highest accuracy of 91.7%, indicating that almost 92% of jobseekers were correctly classified as either unemployed for more than 12 months or not. Additionally, the precision for this category was 94%; thus, for those individuals predicted by the model to have an unemployment duration of more than 12 months, over 94% experienced such a period of unemployment.

Regarding the macro- and micro-average of the CART ensemble model evaluation statistics, we can say that model has a very good prediction accuracy of over 77%. This means that, on average, the predictions of the model on jobseekers' unemployment duration into one of four categories were correct in more than 77% cases. The sensitivity of the model in correctly identifying the right category of unemployment duration averaged over 53% (according to the macro-average) or nearly 55% (according to the micro-average). Compared to a naïve model that randomly predicts one of the four categories with 25% accuracy, this model achieves highly satisfactory results.

The importance of individual predictors in the model is illustrated in Figure 1. This figure displays the frequency of using individual input variables in the nine component models of the boosted CART model. Only those predictors that were used in the models at least seven times are shown in the figure. The remaining predictors were also used in the component models but are not displayed in Figure 1.

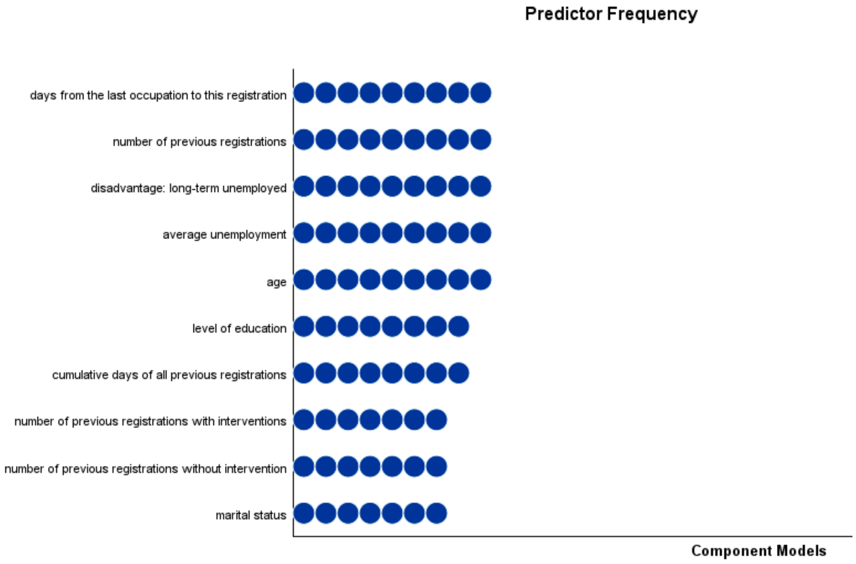


Figure 1. Frequency of using predictors in the component models of the boosted CART model.

We can conclude that the course of previous unemployment plays a crucial role in an individual's future unemployment duration. Among the variables used in all ten component models are, among others, *work before registration* (the number of days from the last occupation to the current registration), *number of previous registrations* and *average unemployment* (average number of days spent in unemployment per year). The variable *cumulative previous registrations* (the cumulative days of previous registrations before the period under review) is missing in only one component model, while *numbers of previous registrations with interventions* or *without interventions* appear in eight out of ten component models.

Since we are unable to illustrate the boosted CART model with a dendrogram, given that it consists of ten component models, we instead present the CART model without the boosting technique in Appendix A, and the same model in a set of rules for the individual values of the target variable in Appendix B. It is very interesting to observe how the input characteristics of jobseekers influence the duration of their unemployment.

3.2. CHAID Model

A CHAID model was also constructed using a boosting technique, emphasising the accuracy of the model. The number of component models was ten, which were then utilised to train the boosted model. The estimated accuracy of the ensemble boosted model was 42.3% in the training set. This can be considered a better accuracy than a naïve model with a 25% prediction accuracy. The coincidence matrix of the CHAID model is presented in Table 7.

Table 7. Coincidence matrix of boosted CHAID model.

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Sum	Correct [%]
up to 3 months	52,351	20,851	22,556	29,267	125,025	0.419
3–6 months	30,143	35,092	22,129	24,668	112,032	0.313
6–12 months	33,378	20,786	42,928	39,584	136,676	0.314
more than 12 months	30,784	16,771	26,399	76,626	150,580	0.509
sum	146,656	93,500	114,012	170,145	524,313	0.395

To evaluate the sensitivity, specificity, precision, and F-measure of the model, we recalculated the confusion matrix into a 4-field matrix for each category of the unemployment duration and then summed these individual tables to compute the macro- and micro-averages. The cumulative 4-field confusion matrix of model is presented in Table 8.

Table 8. Cumulative confusion matrix for boosted CHAID model.

Unemployment Duration	Sum Correct	Sum Incorrect	Sum	Correct [%]
sum correct	206,997	317,316	524,313	**39.48**
sum incorrect	317,316	1,255,623	1,572,939	**79.83**
sum	524,313	1,572,939	2,097,252	**69.74**
correct [%]	**39.48**	20.17		

The cumulative and individual tables were used to calculate the evaluation measures of this model, which we present in Table 9 for all categories (first four columns) of the target variable together with their macro- and micro-averages across the entire CHAID model (last two columns).

Table 9. Evaluation measures for boosted CHAID model.

Evaluation Measure	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Macro-Average	Micro-Average
sensitivity	41.87	31.32	31.41	50.89	38.87	39.48
specificity	76.38	85.83	81.66	74.98	79.71	79.83
precision	35.70	37.53	37.65	45.04	38.98	39.48
F-measure	38.54	34.15	34.25	47.78	38.68	56.60
accuracy	35.70	74.19	68.56	68.06	69.74	69.74

All the mentioned measures were calculated for the testing set to depict the true quality of the model based on the data which were not used during the model's creation. The CHAID model demonstrated, similarly to the CART model, the best performance in identifying the category of jobseekers unemployed for more than 12 months, with a sensitivity of over 50%. The highest accuracy of 74.2% was achieved in the group of jobseekers with unemployment lasting 3 to 6 months. That means that over 74% of jobseekers were correctly classified as either unemployed for this period or not. On the other hand, the model achieved a weaker performance in the category of jobseekers unemployed for up to 3 months but still functioned as a better model than a naïve random prediction of unemployment duration.

The accuracy of almost 70% can describe the overall performance of the CHAID model. This means that, on average, the model correctly predicted the duration of unemployment of jobseekers in almost 70% of cases. The sensitivity of the model is almost 39 (macro-average) to almost 40% (according to the micro-average).

The frequency of the individual predictors used in the component models is illustrated in Figure 2. The figure does not display the variables used less than nine times, although they are also used in the component models.

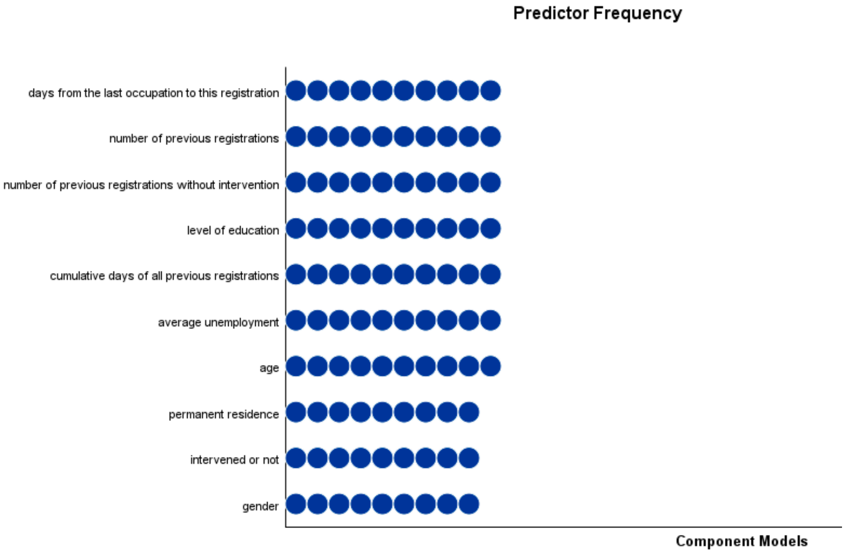


Figure 2. Frequency of using predictors in the component models of the boosted CHAID model.

According to the frequency of the individual variables used in the component models, the history of a jobseeker's previous unemploymentok is very important in their future duration of unemployment registration. Besides the variables describing the unemployment history, the level of education, age, permanent residence, and gender are the most used variables in the component models. The component models contain between 10 and 12 predictors. Considering the complexity of the CHAID ensemble model, we cannot depict it using the dendrogram. We present one component model instead, but not as a dendrogram, as it has 63 nodes altogether, and the picture would be unreadable. In Appendix C, the set of rules for the individual categories of unemployment duration is listed instead.

3.3. Discriminant Model

The discriminant analysis model was constructed as a boosted model for increasing the accuracy of the model and produced predictions of the discriminant score for each officially registered jobseeker. According to this score, a jobseeker is categorised into one of four categories of unemployment duration.

For the quality of the model, it was important to assess the discriminant ability of the input variables that were later used in the process of the model's creation for selecting important variables. It can be checked using the values of the canonical discriminant function coefficients in the standardised form suitable for their comparison. These coefficients are presented in Table 10, together with the correlations between the discriminating variables and discriminant functions (in parentheses).

Table 10. Standardised canonical discriminant function coefficients, and their correlation with discriminant functions.

Function	Up to 3 Months	3–6 Months	6–12 Months
age	−0.086 (−0.135)	0.612 (0.723)	0.149 (−0.062)
work before registration	0.289 (0.113)	0.488 (0.708)	−0.718 (−0.562)
cumulative previous registrations	0.734 (0.723)	0.298 (0.239)	0.610 (0.612)
average unemployment	−0.692 (−0.614)	0.306 (0.460)	0.552 (0.422)

According to this table, the discriminant model utilises only four input variables as significant predictors for unemployment duration. A similar result can be seen in Figure 3.

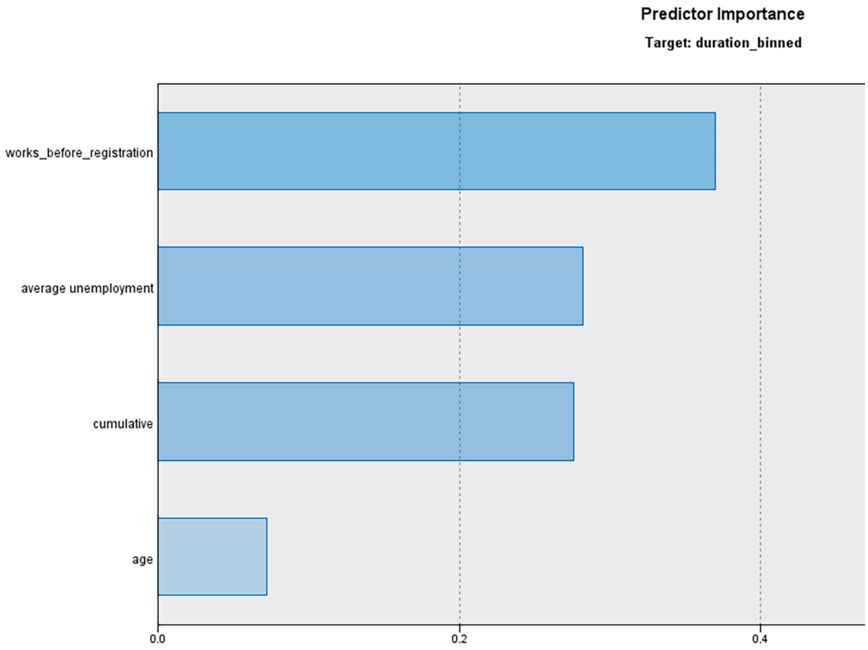


Figure 3. Predictor importance in the discriminant model.

Among these variables, the cumulative unemployment history (in the figure, regarded as *cumulative*) is the best discriminatory variable for the shortest unemployment duration. The best discriminatory variable for 3–6 and 6–12 months of unemployment duration is the number of days from the last occupation to the current registration. This variable is the most important predictor in the whole discriminant model. Overall, according to the discriminant analysis result, unemployment history is very important in the future unemployment duration of jobseekers. Moreover, a jobseeker's age is also very important.

The coincidence matrix of the discriminant model is shown in Table 11.

Table 11. Coincidence matrix of the boosted discriminant model.

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Sum	Correct [%]
up to 3 months	80,095	11,131	7651	26,148	125,025	0.641
3–6 months	58,857	20,914	7879	24,382	112,032	0.187
6–12 months	71,058	15,161	10,756	39,701	136,676	0.079

Table 11. *Cont.*

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Sum	Correct [%]
more than 12 months	71,596	12,336	7016	59,632	150,580	0.396
sum	281,606	59,542	33,302	149,863	524,313	0.327

The cumulative 4-field confusion matrix for this model, recalculated to calculate the averaged evaluation statistics, is presented in Table 12.

Table 12. Cumulative confusion matrix for the boosted discriminant model.

Unemployment Duration	Sum Correct	Sum Incorrect	Sum	Correct [%]
sum correct	171,397	352,916	524,313	**32.69**
sum incorrect	352,916	1,220,023	1,572,939	**77.56**
sum	524,313	1,572,939	2,097,252	**66.34**
correct [%]	**32.69**	22.44		

The evaluation measures for each category of unemployment duration and their macro- and micro-averages for the entire discriminant model are presented in Table 13.

Table 13. Evaluation measures for boosted discriminant model.

Evaluation Measure	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Macro-Average	Micro-Average
sensitivity	64.06	18.67	7.87	39.60	32.55	32.69
specificity	49.53	90.63	94.18	75.86	77.55	77.56
precision	28.44	35.12	32.30	39.79	33.91	32.69
F-measure	39.39	24.38	12.66	39.70	29.03	32.69
accuracy	28.44	75.25	71.68	65.44	66.34	66.34

The evaluation measures listed in Table 13 were calculated for the testing set, thus depicting the real quality of the model. The discriminant model demonstrated its best performance in the category of jobseekers unemployed for up to 3 months, with a sensitivity of over 64%. The highest accuracy of 75.3% was achieved in the group of jobseekers with unemployment lasting from 3 to 6 months. On the other hand, the model achieved a weaker accuracy in the category of jobseekers unemployed for up to 3 months.

The overall accuracy of the discriminant model was more than 66%, i.e., on average, the correctness of the discriminant model in the prediction of the unemployment duration of jobseekers by clustering them into one of four time categories achieved 66% accuracy. The sensitivity of this model was almost 33% according to the macro- and micro-averages.

The discriminant model can be described as a set of discriminant functions for all categories of unemployment duration (Fisher's discriminant functions). Their coefficients are presented in Table 14. A jobseeker is then predicted to be unemployed for a period with the highest value of discriminant functions.

Table 14. Coefficients of discriminant functions.

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months
age	0.227	0.242	0.240	0.240
work before registration	−0.001	0.0004	0.0005	−0.001
cumulative	0.002	0.003	0.002	0.002
average unemployment	0.011	0.012	0.014	0.018
(constant)	−50.520	−60.201	−60.126	−60.119

3.4. Logistic Regression Meta-Model

We created the final model for predicting unemployment duration using the stacking method, incorporating the predictions of individual models (CART, CHAID, and discriminant analysis). These predictions were used as input variables (prefixed with pred-) along with a confidence measure from each model (prefixed with confidence-). This approach enabled us to create a more accurate model than the individual models alone. As observed in the previous models, each model exhibited accuracy in different categories. Hence, we supposed that combining these models could lead to an improved identification of the appropriate unemployment duration category.

The final meta-model was created using logistic regression and predicts not only the category of the unemployment duration for individual jobseekers but also the probabilities of all unemployment categories according to duration. Subsequently, jobseekers are assigned to the unemployment duration category with the highest probability, or predetermined criteria based on these probabilities can be made to categorise the jobseekers. The following results were obtained using the highest probability as the determinant for categorising jobseekers. Table 15 provides details on the logistic regression model and its coefficients, with the reference category being the first category, that is *up to 3 months*.

Table 15. Coefficients of the logistic regression meta-model.

Unemployment Duration	3–6 Months	6–12 Months	More than 12 Months
Intercept	−2.014	0.475	−0.686
pred-CART_cat1	0.102	−0.738	−1.745
pred-CART_cat2	0.870	−0.390	−2.266
pred-CART_cat3	0.425	−0.023	−1.361
confidence-CART	0.759	−2.932	3.768
pred-CHAID _cat1	−0.074	−0.451	−2.591
pred-CHAID _cat2	0.602	−0.054	−1.861
pred-CHAID _cat3	0.380	0.336	−1.820
confidence-CHAID	2.379	2.375	3.087
pred-discriminant_cat1	−0.035	−0.158	−0.263
pred-discriminant_cat2	0.144	−0.036	−0.453
pred-discriminant_cat3	0.046	0.196	−0.096
confidence-discriminant	0.474	0.414	0.971

This model utilises the predictions and their confidence from the abovementioned models. Their importance in the final meta-model is depicted in Figure 4.

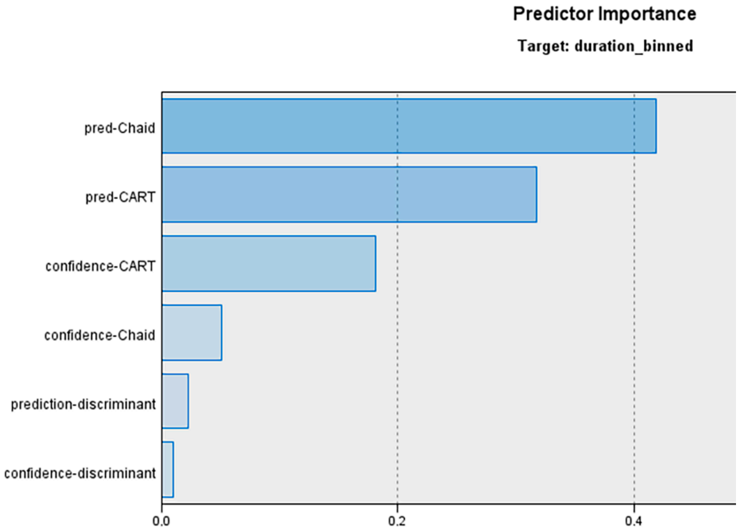


Figure 4. Predictor importance in the final meta-model.

Finally, we evaluated the final meta-model of logistic regression using the evaluation measures. The coincidence matrix for the testing set is shown in Table 16.

Table 16. Coincidence matrix of final meta-model.

Unemployment Duration	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Sum	Correct [%]
up to 3 months	68,854	18,414	34,376	3381	125,025	0.551
3–6 months	40,017	33,368	35,863	2784	112,032	0.298
6–12 months	44,096	19,257	68,413	4910	136,676	0.501
more than 12 months	11,411	4467	15,962	118,740	150,580	0.789
sum	164,378	75,506	154,614	129,815	524,313	0.552

The cumulative 4-field confusion matrix for this model is presented in Table 17. The rows show real situations, and the columns are the predictions of the unemployment duration category from the model.

Table 17. Cumulative confusion matrix for final meta-model.

Unemployment Duration	Sum Correct	Sum Incorrect	Sum	Correct [%]
sum correct	289,375	234,938	524,313	**55.19**
sum incorrect	234,938	1,338,001	1,572,939	**85.06**
sum	524,313	1,572,939	2,097,252	**77.60**
correct [%]	**55.19**	14.94		

The evaluation measures for each category of unemployment duration and their macro- and micro-averages for the entire final meta-model are presented in Table 18.

Table 18. Evaluation measures for the final meta-model.

Evaluation Measure	Up to 3 Months	3–6 Months	6–12 Months	More than 12 Months	Macro-Average	Micro-Average
sensitivity	55.07	29.78	50.05	78.86	53.44	55.19
specificity	76.08	89.78	77.76	97.04	85.16	85.06
precision	41.89	44.19	44.25	91.47	55.45	55.19
F-measure	47.58	35.59	46.97	84.69	53.71	55.19
accuracy	41.89	76.96	70.54	91.82	77.60	77.60

The final model achieved an average accuracy of 77.60% (both macro- and micro-average), indicating that the predictions obtained from the model were correct in almost 78% of cases. In the category of jobseekers who were potentially unemployed for more than 12 months, the model demonstrated a high accuracy of almost 92% and more than 91% precision. This implies that our model can accurately identify problematic jobseekers at risk of long-term unemployment. When the model predicted that a jobseeker would be unemployed for more than 12 months, this prediction was correct in more than 91% of cases.

Moreover, the model also achieved high accuracy for jobseekers unemployed for 3–6 months and 6–12 months. Thus, if the model predicts these unemployment durations, it is likely to be correct in most cases. The sensitivity of the overall model in correctly identifying the appropriate category of unemployment duration was over 53% (macro-average) or over 55% (micro-average).

4. Discussion

When creating the CART model, the CHAID model, and the discriminant model, the history of the previous unemployment of the jobseeker, which was represented by variables such as the number of days from the last occupation to the current registration, the number of previous registrations, and the average number of days spent in unemployment per year, was found to be the most important factor in predicting unemployment duration. This is essentially related to the findings of a study [61] where the authors sent fictitious resumes to job offers in the US to see if unemployment debt affects whether a potential employer will contact them for an interview. And they found that the probability of being invited to a job interview decreases as the period of unemployment increases.

In the case of the created CHAID model, whose ability to predict the correct classification into one of the four groups of the duration of unemployment of jobseekers is almost 70% accurate, the level of education and age also proved to be important predictors in addition to the unemployment history of the jobseeker, which was in the ten component models of the boosted CHAID model. Similarly, in [17], the author found that age negatively affects the probability of employment, which can be explained by the fact that older workers are less adaptable to changing labour market conditions. On the contrary, the level of education achieved by a jobseeker positively affects their probability of employment, which means that people with higher education can look for work more effectively. These findings are in accordance with the findings of the study [62] on the conditions of the Slovak Republic, where the authors confirmed that the probability of employment is higher for persons with a lower age and higher education. In [63], the author developed a simulator to predict the duration of unemployment of jobseekers. Based on the results of the MCA analysis that was carried out, the area of permanent residence and age appear to be significant variables influencing the duration of unemployment of jobseekers. When creating a multinomial regression model, not only the area of permanent residence and age but also gender and level of education or diploma appear as statistically significant variables that impact the duration of unemployment of jobseekers.

The logistic regression meta-model created by us can predict the category of the length of unemployment for individual job seekers with an accuracy of almost 78%. It distinguishes four categories of unemployment duration: up to 3 months, 3–6 months, 6–12 months, and more than 12 months. In the study by [64], the authors use a logistic regression model to include an individual in one of two categories of unemployment duration, namely short-term (up to 12 months) and long-term unemployment (over 12 months), during a period of economic boom and fall in Estonia. The models created in this study, based on the Nagelkerke R Square value, described the variability of the dependent variable in the range of only 0.001–0.044.

5. Conclusions

In this study, we aimed to predict the duration of unemployment of jobseekers in Slovakia using an ensemble model created through the stacking method. Our approach involved developing three models using CART, CHAID, and discriminant analysis techniques. These models provided predictions for unemployment duration, and their outputs and confidence were used as input variables for a final meta-model created via logistic regression.

If we were to examine the predictive ability of certain models in greater detail, the first model we developed was the CART model. Using a boosting technique, the CART model was constructed to improve its accuracy. Based on the macro- and micro-averages of the CART model evaluation statistics, we can conclude that the model's prediction accuracy exceeded 77%. This indicates that the given model accurately placed unemployed job seekers into one of the four categories of unemployment duration with an accuracy of greater than 77%. This model had the highest sensitivity (up to 75.84%) for identifying the category of long-term unemployed job seekers (those unemployed for more than 12 months). The CHAID model was the second model created. This model's accuracy in predicting the duration of unemployment of jobseekers was slightly lower than the previous model, at approximately 70%. Thus, almost 70% of the time, the model's predictions regarding the length of unemployment for job seekers in one of the four categories were accurate. The CHAID model performed best in identifying the category of job seekers who had been unemployed for over a year, with a sensitivity of over 50%. The last model we developed was the discriminant model. This model's overall accuracy was greater than 66%. The average accuracy of the model's predictions regarding the duration of unemployment of job seekers in one of the four categories was 66%. With a sensitivity of over 64%, this model demonstrated the best performance in the category of unemployment duration of unemployed job seekers within three months.

Our ensemble model for predicting unemployment duration achieved an overall accuracy of 77.60%. The final model classified unemployed job seekers into one of four groups of duration of unemployment: within three months, between three and six months, between six and twelve months, and more than twelve months. For the category of unemployment up to 3 months, the model's accuracy was nearly 42%; for the category of unemployment from 3 to 6 months, the model's accuracy exceeded 76%; and for the category of unemployment from 6 to 12 months, the model's accuracy was 70.54%. Notably, it exhibited exceptional performance in predicting the unemployment duration of jobseekers who remained unemployed for more than 12 months, with an accuracy of nearly 92% and a precision exceeding 91%. This indicates the potential of our model to identify those jobseekers threatened with long-term unemployment accurately.

In addition, when examining the significance of the individual predictors of unemployment duration in the individual models, unemployment history, level of education, and age appear to be important. Our findings highlight the significance of a jobseeker's unemployment history in determining their future unemployment duration. This implies that prior unemployment experiences can serve as valuable indicators for forecasting the length of unemployment of jobseekers. The performance of our ensemble model emphasises the potential practical applications of our research. As the ensemble model was shown to be very accurate at predicting long-term unemployment (over 12 months), it has the potential to aid employment offices in identifying unemployed individuals who may be at a greater risk of long-term unemployment. By accurately predicting unemployment durations, policymakers and employment agencies can devise targeted interventions and support systems to assist jobseekers in finding employment more effectively.

The continuation of this study could be aimed at verifying the prediction ability of the created ensemble model during the following period. For this purpose, it will be necessary to obtain new data from the Ministry of Labour, Social Affairs and Family of the Slovak Republic. It would be appropriate to verify the functionality of the model in a non-standard situation, such as the one brought about by the COVID-19 pandemic, which had a strong impact on unemployment. If it is possible due to the availability of data, we also intend to verify the functionality of the model on data from another country, after taking the national specifics into account.

In addition, we want to focus on the application of several machine learning methods to strengthen the predictive ability of the created model in those categories of unemployment duration where its strength was lower.

Author Contributions: Conceptualisation, B.G. and L.S.; methodology, L.S.; software, L.S.; validation, B.G. and K.K.; formal analysis, K.K.; investigation, B.G.; resources, K.K.; data curation, B.G. and L.S.; writing—original draft preparation, B.G. and L.S.; writing—review and editing, B.G., L.S. and K.K.; visualisation, L.S.; supervision, L.S.; project administration, K.K.; funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This publication was created thanks to support under the Operational Program Integrated Infrastructure for the project "The implementation framework and business model of the Internet of Things, Industry 4.0 and smart transport." (ITMS code: 313011BWN6), co-financed by the European Regional Development Fund.

Institutional Review Board Statement: Not applicable.

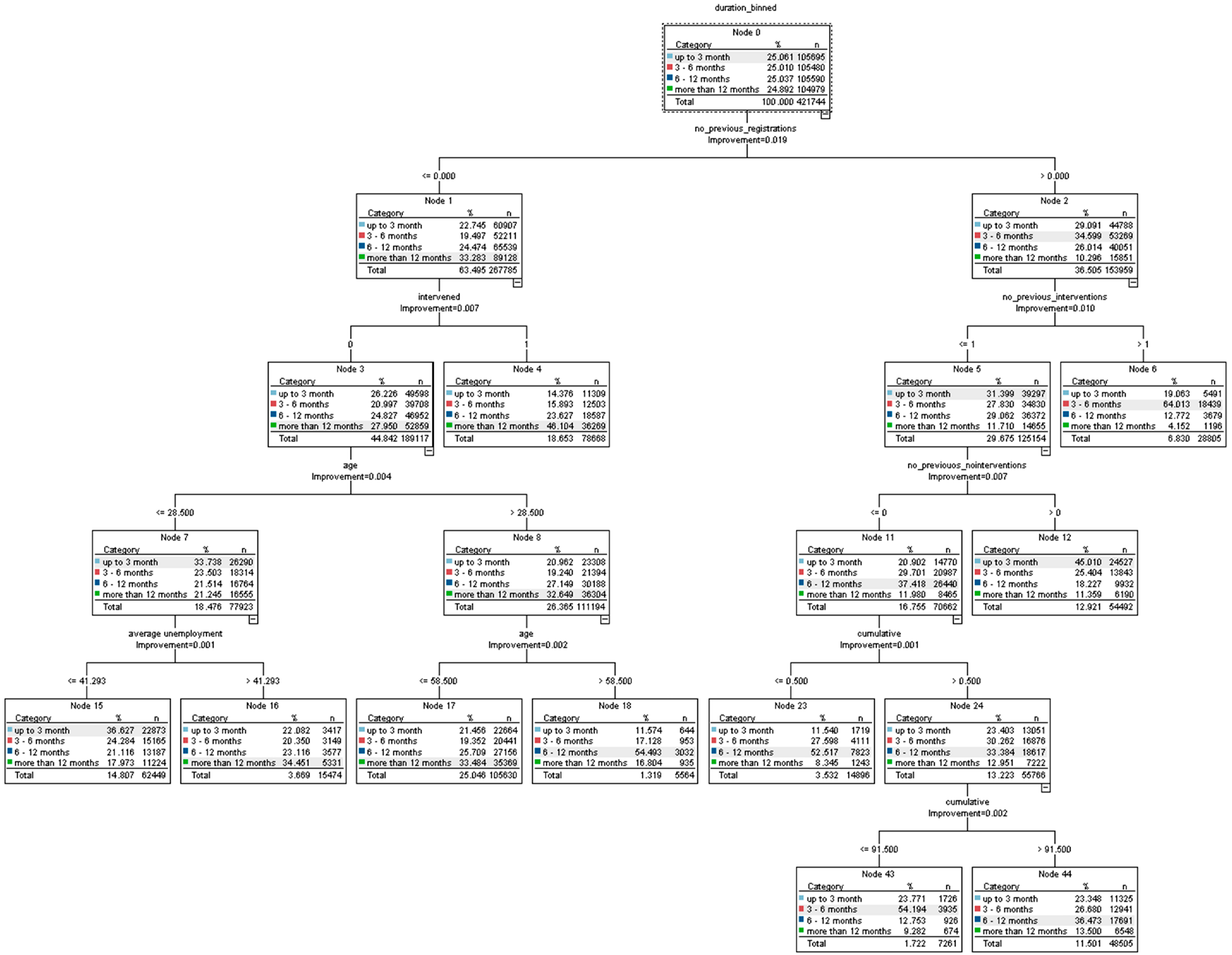
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the Ministry of Labour, Social Affairs and Family of the Slovak Republic, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission from the Ministry of Labour, Social Affairs and Family of the Slovak Republic.

Acknowledgments: This article has benefited from constructive feedback on earlier drafts from the editors, the chief editor, and the anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Dendrogram of the CART Model


duration_binned
Node 0
Category % n
🟦 up to 3 month 25.061 105695
🟥 3 - 6 months 25.010 105480
🟦 6 - 12 months 25.037 105590
🟩 more than 12 months 24.892 104979
Total 100.000 421744
no_previous_registrations
Improvement=0.019
<= 0.000
Node 1
Category % n
up to 3 month 22.745 60907
3 - 6 months 19.497 52211
6 - 12 months 24.474 65539
more than 12 months 33.283 89128
Total 63.495 267785
> 0.000
Node 2
Category % n
up to 3 month 29.091 44788
3 - 6 months 34.599 53269
6 - 12 months 26.014 40051
more than 12 months 10.296 15851
Total 36.505 153959
intervened
Improvement=0.007
0
Node 3
Category % n
up to 3 month 26.226 49598
3 - 6 months 20.997 39708
6 - 12 months 24.827 46952
more than 12 months 27.950 52859
Total 44.842 189117
1
Node 4
Category % n
up to 3 month 14.376 11309
3 - 6 months 15.893 12503
6 - 12 months 23.627 18587
more than 12 months 46.104 36269
Total 18.653 78668
no_previous_interventions
Improvement=0.010
<= 1
Node 5
Category % n
up to 3 month 31.399 39297
3 - 6 months 27.830 34830
6 - 12 months 29.062 36372
more than 12 months 11.710 14655
Total 29.675 125154
> 1
Node 6
Category % n
up to 3 month 19.063 5491
3 - 6 months 64.013 18439
6 - 12 months 12.772 3679
more than 12 months 4.152 1196
Total 6.830 28805
age
Improvement=0.004
<= 28.500
Node 7
Category % n
up to 3 month 33.738 26290
3 - 6 months 23.503 18314
6 - 12 months 21.514 16764
more than 12 months 21.245 16555
Total 18.476 77923
> 28.500
Node 8
Category % n
up to 3 month 20.962 23308
3 - 6 months 19.240 21394
6 - 12 months 27.149 30188
more than 12 months 32.649 36304
Total 26.365 111194
no_previouos_nointerventions
Improvement=0.007
<= 0
Node 11
Category % n
up to 3 month 20.902 14770
3 - 6 months 29.701 20987
6 - 12 months 37.418 26440
more than 12 months 11.980 8465
Total 16.755 70662
> 0
Node 12
Category % n
up to 3 month 45.010 24527
3 - 6 months 25.404 13843
6 - 12 months 18.227 9932
more than 12 months 11.359 6190
Total 12.921 54492
average unemployment
Improvement=0.001
<= 41.293
Node 15
Category % n
up to 3 month 36.627 22873
3 - 6 months 24.284 15165
6 - 12 months 21.116 13187
more than 12 months 17.973 11224
Total 14.807 62449
> 41.293
Node 16
Category % n
up to 3 month 22.082 3417
3 - 6 months 20.350 3149
6 - 12 months 23.116 3577
more than 12 months 34.451 5331
Total 3.669 15474
age
Improvement=0.002
<= 58.500
Node 17
Category % n
up to 3 month 21.456 22664
3 - 6 months 19.352 20441
6 - 12 months 25.709 27156
more than 12 months 33.484 35369
Total 25.046 105630
> 58.500
Node 18
Category % n
up to 3 month 11.574 644
3 - 6 months 17.128 953
6 - 12 months 54.493 3032
more than 12 months 16.804 935
Total 1.319 5564
cumulative
Improvement=0.001
<= 0.500
Node 23
Category % n
up to 3 month 11.540 1719
3 - 6 months 27.598 4111
6 - 12 months 52.517 7823
more than 12 months 8.345 1243
Total 3.532 14896
> 0.500
Node 24
Category % n
up to 3 month 23.403 13051
3 - 6 months 30.262 16876
6 - 12 months 33.384 18617
more than 12 months 12.951 7222
Total 13.223 55766
cumulative
Improvement=0.002
<= 91.500
Node 43
Category % n
up to 3 month 23.771 1726
3 - 6 months 54.194 3935
6 - 12 months 12.753 926
more than 12 months 9.282 674
Total 1.722 7261
> 91.500
Node 44
Category % n
up to 3 month 23.348 11325
3 - 6 months 26.680 12941
6 - 12 months 36.473 17691
more than 12 months 13.500 6548
Total 11.501 48505

Appendix B. CART Model in Rules

Rules for cat1—contains two rules

Rule 1: if age ≤ 28.5 and average unemployment ≤ 37.774 and n_previous_registrations = 0 and intervened = 0 then cat1

Rule 2: if n_previouos_nointerventions ≤ 14 and n_previous_registrations ≤ 14 and n_previous_interventions in ≤ 1 then cat1

Rules for cat2—contains two rules

Rule 1: if cumulative > 0.5 and cumulative ≤ 91.5 and n_previouos_nointerventions = 0 and n_previous_registrations ≤ 14 and n_previous_interventions ≤ 1 then cat2

Rule 2: if n_previous_registrations ≤ 14 and n_previous_interventions in [2 3 4 5 6 7 8 9 10 11 12] then cat2

Rules for cat3—contains three rules

Rule 1: if age > 28.5 and age > 58.5 and n_previous_registrations = 0 and intervened = 0 then cat3

Rule 2: if cumulative ≤ 0.5 and n_previouos_nointerventions = 0 and n_previous_registrations ≤ 14 and n_previous_interventions ≤ 1 then cat3

Rule 3: if cumulative > 91.500 and n_previouos_nointerventions = 0 and n_previous_registrations ≤ 14 and n_previous_interventions ≤ 1 then cat3

Rules for cat4—contains three rules

Rule 1: if age ≤ 28.5 and average unemployment > 37.774 and n_previous_registrations = 0 and intervened = 0 then cat4

Rule 2: if age > 28.5 and age ≤ 58.5 and n_previous_registrations = 0 and intervened = 0 then cat4

Rule 3: if n_previous_registrations = 0 and intervened = 1 then cat4

Default: cat4

Appendix C. CHAID Model in Rules

Rules for cat1—contains 21 rules

Rule 1: if intervened = 0 and age ≤ 20 and education ≤ 2 then cat1

Rule 2: if intervened = 0 and age ≤ 20 and education > 2 and education ≤ 4 then cat1

Rule 3: if intervened = 0 and age ≤ 20 and education > 4 and cumulative ≤ 183 then cat1

Rule 4: if intervened = 0 and age ≤ 20 and education > 4 and cumulative > 183 then cat1

Rule 5: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment ≤ 7.593 and education ≤ 6 then cat1

Rule 6: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment ≤ 7.593 and education > 6 then cat1

Rule 7: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment > 7.593 and average unemployment ≤ 25.657 then cat1

Rule 8: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment > 25.657 and average unemployment ≤ 77.935 and n_previouos_nointerventions = 0 then cat1

Rule 9: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment > 25.657 and average unemployment ≤ 77.935 and n_previouos_nointerventions > 0 then cat1

Rule 10: if intervened = 0 and age > 20 and age ≤ 24 and average unemployment > 77.935 then cat1

Rule 11: if intervened = 0 and age > 24 and age ≤ 28 and average unemployment ≤ 15.711 and education ≤ 7 then cat1

Rule 12: if intervened = 0 and age > 24 and age ≤ 28 and average unemployment ≤ 15.711 and education > 7 then cat1

Rule 13: if intervened = 0 and age > 24 and age ≤ 28 and average unemployment > 15.711 and average unemployment ≤ 37.968 then cat1

Rule 14: if intervened = 0 and age > 24 and age ≤ 28 and average unemployment > 37.968 and cumulative > 0 then cat1

Rule 15: if intervened = 0 and age > 28 and age ≤ 32 and n_previouos_nointerventions ≤ 0 and average unemployment ≤ 37.968 then cat1

Rule 16: if and intervened = 0 and age > 28 and age ≤ 32 and n_previouos_nointerventions > 0 then cat1

Rule 17: if intervened = 0 and age > 32 and age ≤ 42 and n_previouos_nointerventions > 0 and cumulative ≤ 183 then cat1

Rule 18: if intervened = 0 and age > 32 and age ≤ 42 and n_previouos_nointerventions > 0 and cumulative > 183 then cat1

Rule 19: if intervened = 0 and age > 42 and age ≤ 48 and n_previouos_nointerventions > 0 then cat1

Rule 20: if and intervened = 0 and age > 48 and age ≤ 54 and n_previous_registrations > 0 then cat1

Rule 21: if intervened = 1 and n_previous_interventions ≤ 0 and works_before_registration > 361 and works_before_registration ≤ 1 499 then cat1

Rules for cat2—contains five rules
Rule 1: if intervened = 1 and n_previous_interventions ≤ 1 and works_before_registration ≤ 19 then cat2
Rule 2: if intervened = 1 and n_previous_interventions ≤ 1 and works_before_registration > 19 then cat2
Rule 3: if intervened = 1 and n_previous_interventions > 1 and n_previous_interventions ≤ 2 then cat2
Rule 4: if intervened = 1 and n_previous_interventions > 2 and n_previous_interventions ≤ 4 then cat2
Rule 5: if intervened = 1 and n_previous_interventions > 4 then cat2
Rules for cat3—contains 25 rules
Rule 1: if dis3 = 1 and n_previous_registrations ≤ 1 and works_before_registration = 0 and marital_status = 0 or 1 then cat3
Rule 2: if dis3 = 1 and n_previous_registrations ≤ 1 and works_before_registration = 0 and marital_status = 2 or 3 or 4 then cat3
Rule 3: if dis3 = 1 and n_previous_registrations ≤ 1 and works_before_registration > 361 and works_before_registration ≤ 1 499 then cat3
Rule 4: if dis3 = 1 and n_previous_registrations ≤ 1 and works_before_registration > 1 499 then cat3
Rule 5: if dis3 = 1 and n_previous_registrations > 1 and works_before_registration ≤ 1 then cat3
Rule 6: if dis3 = 1 and n_previous_registrations > 1 and works_before_registration > 1 then cat3
Rule 7: if intervened = 0 and age > 24 and age ≤ 28 and average unemployment > 37.968 and cumulative ≤ 0 then cat3
Rule 8: if intervened = 0 and age > 28 and age ≤ 32 and n_previouos_nointerventions ≤ 0 and average unemployment > 37.968 then cat3
Rule 9: if and intervened = 0 and age > 32 and age ≤ 42 and n_previouos_nointerventions ≤ 0 and average unemployment ≤ 25.657 then cat3
Rule 10: if intervened = 0 and age > 32 and age ≤ 42 and n_previouos_nointerventions ≤ 0 and average unemployment > 25.657 and average unemployment ≤ 77.935 then cat3
Rule 11: if intervened = 0 and age > 32 and age ≤ 42 and n_previouos_nointerventions ≤ 0 and average unemployment > 77.935 then cat3
Rule 12: if intervened = 0 and age > 42 and age ≤ 48 and n_previouos_nointerventions ≤ 0 and average unemployment ≤ 15.711 then cat3
Rule 13: if intervened = 0 and age > 42 and age ≤ 48 and n_previouos_nointerventions ≤ 0 and average unemployment > 15.711 then cat3
Rule 14: if intervened = 0 and age > 48 and age ≤ 54 and n_previous_registrations ≤ 0 and gender = 1 then cat3
Rule 15: if intervened = 0 and age > 48 and age ≤ 54 and n_previous_registrations ≤ 0 and gender = 2 then cat3
Rule 16: if intervened = 0 and age > 54 and average unemployment ≤ 15.711 then cat3
Rule 17: if intervened = 0 and age > 54 and average unemployment > 15.711 then cat3
Rule 18: if intervened = 1 and n_previous_interventions = 0 and works_before_registration ≤ 0 and age ≤ 20 then cat3
Rule 19: if and intervened = 1 and n_previous_interventions = 0 and works_before_registration ≤ 0 and age > 20 and age ≤ 28 then cat3
Rule 20: if intervened = 1 and n_previous_interventions = 0 and works_before_registration = 0 and age > 28 then cat3
Rule 21: if intervened = 1 and n_previous_interventions = 0 and works_before_registration ≤ 19 and age ≤ 28 then cat3
Rule 22: if intervened = 1 and n_previous_interventions = 0 and works_before_registration ≤ 19 and age > 28 and age ≤ 42 then cat3
Rule 23: if intervened = 1 and n_previous_interventions = 0 and works_before_registration ≤ 19 and age > 42 then cat3
Rule 24: if intervened = 1 and n_previous_interventions = 0 and works_before_registration > 19 and works_before_registration ≤ 361 then cat3
Rule 25: if intervened = 1 and n_previous_interventions = 0 and works_before_registration > 1 499 then cat3
Rules for cat4—contains 12 rules
Rule 1: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age ≤ 24 then cat4
Rule 2: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age > 24 and age ≤ 54 and works_before_registration = 0 then cat4
Rule 3: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age > 24 and age ≤ 54 and and works_before_registration ≤ 1 then cat4
Rule 4: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age > 24 and age ≤ 54 and works_before_registration > 1 and works_before_registration ≤ 361 then cat4

Rule 5: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age > 24 and age $\leq$ 54 and works_before_registration > 361 then cat4
Rule 6: if dis3 = 1 and n_previous_registrations = 0 and intervened = 0 and age > 54 then cat4
Rule 7: if dis3 = 1 and n_previous_registrations = 0 and intervened = 1 and works_before_registration = 0 and gender = 1 then cat4
Rule 8: if dis3 = 1 and n_previous_registrations = 0 and intervened = 1 and works_before_registration = 0 and gender = 2 then cat4
Rule 9: if dis3 = 1 and n_previous_registrations = 0 and intervened = 1 and works_before_registration $\leq$ 361 and marital_status = 1 or marital_status = 4 then cat4
Rule 10: if dis3 = 1 and n_previous_registrations = 0 and intervened = 1 and works_before_registration $\leq$ 361 and marital_status = 2 or marital_status = 3 then cat4
Rule 11: if dis3 = 1 and n_previous_registrations = 0 and intervened = 1 and works_before_registration > 361 then cat4
Rule 12: if dis3 = 1 and n_previous_registrations = 1 and works_before_registration $\leq$ 361 then cat4
Default: cat3

References

1. Achdut, N.; Refaeli, T. Unemployment and Psychological Distress among Young People during the COVID-19 Pandemic: Psychological Resources and Risk Factors. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7163. [CrossRef] [PubMed]
2. Bennett, P.; Ouazad, A. Job Displacement, Unemployment, and Crime: Evidence from Danish Microdata and Reforms. *J. Eur. Econ. Assoc.* **2020**, *18*, 2182–2220. [CrossRef]
3. Calmfors, L. Labour Market Policy and Unemployment. *Eur. Econ. Rev.* **1995**, *39*, 583–592. [CrossRef]
4. Baliak, M.; Belin, M. *The Current State of Unemployment and Its Short-Term Forecast [Aktualny Stav Nezamestnanosti a Jej Kratkodoba Prognoza]*; Institute of Social Policy: Bratislava, Slovakia, 2020.
5. Barcakova, M.; Janas, K. Youth Unemployment in Slovakia and in Slovenia. *Izzivi Prihodnosti Chall. Future* **2019**, *4*, 98–105.
6. Caliendo, M.; Schmidl, R. Youth Unemployment and Active Labor Market Policies in Europe. *IZA J. Labor Policy* **2016**, *5*, 1. [CrossRef]
7. Banociova, A.; Martinkova, S. Active Labour Market Policies of Selected European Countries and Their Competitiveness. *J. Compet.* **2017**, *9*, 5–21. [CrossRef]
8. Card, D.; Kluve, J.; Weber, A. What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *J. Eur. Econ. Assoc.* **2018**, *16*, 894–931. [CrossRef]
9. Katris, C. Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Comput. Econ.* **2020**, *55*, 673–706. [CrossRef]
10. Viljanen, M.; Pahikkala, T. Predicting Unemployment with Machine Learning Based on Registry Data. In *Research Challenges in Information Science*; Springer International Publishing: Berlin, Germany, 2020; pp. 352–368. ISBN 978-3-030-50315-4.
11. Niyadurupola, V.; Esposito, L. What Gets Them Going? The Effects of Activation Policies on Personal Change Processes of Unemployed Youth. *J. Educ. Work* **2021**, *34*, 590–609. [CrossRef]
12. Prasasti, N.; Ohwada, H. Applicability of Machine-Learning Techniques in Predicting Customer Defection. In Proceedings of the 2014 International Symposium on Technology Management and Emerging Technologies, Bandung, Indonesia, 27–29 May 2014; pp. 157–162.
13. Shinde, P.P.; Shah, S. A Review of Machine Learning and Deep Learning Applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–6.
14. Kreiner, A.; Duca, J. Can Machine Learning on Economic Data Better Forecast the Unemployment Rate? *Appl. Econ. Lett.* **2020**, *27*, 1434–1437. [CrossRef]
15. Werken, L.; Smit, V. Exploring the Use of Recurrent Neural Networks for Predicting Inflation and Unemployment. 2019. Available online: https://www.researchgate.net/profile/Victoria-Smit/publication/337171301_Exploring_the_Use_of_Recurrent_Neural_Networks_for_predicting_Inflation_and_Unemployment/links/5dc9bdfa299bf1a47b2ff69c/Exploring-the-Use-of-Recurrent-Neural-Networks-for-predicting-Inflation-and-Unemployment.pdf (accessed on 1 May 2023).
16. Liu, X.; Li, L. Prediction of Labor Unemployment Based on Time Series Model and Neural Network Model. *Comput. Intell. Neurosci.* **2022**, *2022*, e7019078. [CrossRef] [PubMed]
17. Kupets, O. Determinants of Unemployment Duration in Ukraine. *J. Comp. Econ.* **2006**, *34*, 228–247. [CrossRef]
18. Arslan, H.; Senturk, I. Individual Determinants of Unemployment Duration in Turkey. 2018. Available online: https://hdl.handle.net/20.500.12881/3649 (accessed on 1 May 2023).
19. Niragire, F.; Nshimyiryo, A. Determinants of Increasing Duration of First Unemployment among First Degree Holders in Rwanda: A Logistic Regression Analysis. *J. Educ. Work.* **2017**, *30*, 235–248. [CrossRef]
20. Lim, H.-E. Predicting Low Employability Graduates: The Case of Universiti Utara Malaysia. *Singap. Econ. Rev.* **2010**, *55*, 523–535. [CrossRef]

21. Bayrak, R.; Tatli, H. The Determinants of Youth Unemployment: A Panel Data Analysis of OECD Countries. *Eur. J. Comp. Econ.* **2018**, *15*, 231–248. [CrossRef]
22. Logarusic, M.; Kristic, I.R. Determinants of Unemployment in the European Union. *Ekon. Pregl.* **2019**, *70*, 575–602.
23. Bal-Domańska, B. The Impact of Macroeconomic and Structural Factors on the Unemployment of Young Women and Men. *Econ. Change Restruct.* **2022**, *55*, 1141–1172. [CrossRef]
24. Gogas, P.; Papadimitriou, T.; Sofianos, E. Forecasting Unemployment in the Euro Area with Machine Learning. *J. Forecast.* **2022**, *41*, 551–566. [CrossRef]
25. Gong, J.; Lee, C.-T. Research on Unemployment Rate Based on Machine Learning Method: A Case Study of United States from 1976 to 1986. In Proceedings of the International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2022), Huzhou, China, 15–17 April 2022; Volume 12330, pp. 282–296.
26. Kaya, C.; Bishop, M.; Torres, A. The Impact of Work Incentives Benefits Counseling on Employment Outcomes: A National Vocational Rehabilitation Study. *J. Occup. Rehabil.* **2023**. [CrossRef]
27. McMahon, B.T.; Hurley, J.E.; Chan, F.; Rumrill, P.D.; Roessler, R. Drivers of Hiring Discrimination for Individuals with Disabilities. *J. Occup. Rehabil.* **2008**, *18*, 133–139. [CrossRef]
28. Ho, T.-W. Forecasting Unemployment via Machine Learning: The Use of Average Windows Forecasts. 2022. Available online: https://ssrn.com/abstract=3496138 (accessed on 1 July 2023).
29. Papik, M.; Mihalova, P.; Papikova, L. Determinants of Youth Unemployment Rate: Case of Slovakia. *Equilibrium. Q. J. Econ. Econ. Policy* **2022**, *17*, 391–414. [CrossRef]
30. Karsay, A. Structural and Cyclical Drivers of Unemployment Rate. *NBS Work. Pap.* **2020**, *1*, 1–12.
31. Rublikova, E.; Lubyova, M. Estimating ARIMA–ARCH Model Rate of Unemployment in Slovakia. *Progn. Pract.* **2013**, *5*, 275–289.
32. Maas, B. Short-Term Forecasting of the US Unemployment Rate. *J. Forecast.* **2020**, *39*, 394–411. [CrossRef]
33. Vicente, M.R.; Lopez-Menendez, A.J.; Perez, R. Forecasting Unemployment with Internet Search Data: Does It Help to Improve Predictions When Job Destruction Is Skyrocketing? *Technol. Forecast. Soc. Change* **2015**, *92*, 132–139. [CrossRef]
34. Yi, D.; Ning, S.; Chang, C.-J.; Kou, S.C. Forecasting Unemployment Using Internet Search Data via PRISM. *J. Am. Stat. Assoc.* **2021**, *116*, 1662–1673. [CrossRef]
35. Ganaie, M.A.; Hu, M.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
36. Parker, W.S. Ensemble Modeling, Uncertainty and Robust Predictions. *WIREs Clim. Change* **2013**, *4*, 213–223. [CrossRef]
37. Kotsiantis, S.B.; Pintelas, P.E. Combining Bagging and Boosting. *Int. J. Math. Comput. Sci.* **2007**, *1*, 372–381.
38. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
39. Adamko, P.; Siekelova, A. An Ensemble Model for Prediction of Crisis in Slovak Companies. In Proceedings of the 17th International Scientific Conference Globalization and Its Socio-Economic Consequences: Proceedings, Rajecke Teplice, Slovakia, 4–5 October 2017; University of Zilina: Zilina, Slovakia, 2017. Part I. pp. 1–7.
40. Kim, M.-J.; Kang, D.-K. Ensemble with Neural Networks for Bankruptcy Prediction. *Expert Syst. Appl.* **2010**, *37*, 3373–3379. [CrossRef]
41. Pavlicko, M.; Durica, M.; Mazanec, J. Ensemble Model of the Financial Distress Prediction in Visegrad Group Countries. *Mathematics* **2021**, *9*, 1886. [CrossRef]
42. Bhagia, D. Duration Dependence and Heterogeneity: Learning from Early Notice of Layoff. *arXiv* **2023**, arXiv:2305.17344.
43. Mueller, A.I.; Spinnewijn, J. The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection. *IZA Discuss. Pap. Ser.* **2023**, *15955*, 1–41.
44. Song, Y.; Lu, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135. [CrossRef] [PubMed]
45. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; The Wadsworth statistics/probability series; Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, CA, USA, 1984; ISBN 978-1-351-46048-4.
46. Rutkowski, L.; Jaworski, M.; Pietruczuk, L.; Duda, P. The CART Decision Tree for Mining Data Streams. *Inf. Sci.* **2014**, *266*, 1–15. [CrossRef]
47. Priyam, A.; Gupta, R.; Rathee, A.; Srivastava, S. Comparative Analysis of Decision Tree Classification Algorithms. *Int. J. Curr. Eng. Technol.* **2013**, *3*, 334–337.
48. Ozcan, M.; Peker, S. A Classification and Regression Tree Algorithm for Heart Disease Modeling and Prediction. *Healthc. Anal.* **2023**, *3*, 100130. [CrossRef]
49. Kass, G.V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *J. R. Stat. Society. Ser. C Appl. Stat.* **1980**, *29*, 119–127. [CrossRef]
50. Milanovic, M.; Stamenkovic, M. CHAID Decision Tree: Methodological Frame and Application. *Econ. Themes* **2016**, *54*, 563–586. [CrossRef]
51. Ritschard, G. CHAID and Earlier Supervised Tree Methods. In *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences*; McArdle, J.J., Ritschard, G., Eds.; Routeledge: New York, NY, USA, 2013; pp. 48–74.
52. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; John Wiley & Sons: Hoboken, NJ, USA, 2004; ISBN 978-0-471-69115-0.
53. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]

54. Bickel, P.J.; Levina, E. Some Theory for Fisher's Linear Discriminant Function, "Naive Bayes", and Some Alternatives When There Are Many More Variables than Observations. *Bernoulli* **2004**, *10*, 989–1010. [CrossRef]
55. Tabachnick, B.G.; Fidell, L.S.; Ullman, J.B. *Using Multivariate Statistic*, 7th ed.; Pearson: New York, NY, USA, 2019; ISBN 978-0-13-479054-1.
56. Gajdosikova, D.; Lăzăroiu, G.; Valaskova, K. How Particular Firm-Specific Features Influence Corporate Debt Level: A Case Study of Slovak Enterprises. *Axioms* **2023**, *12*, 183. [CrossRef]
57. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: New York, NY, USA, 2013; ISBN 978-0-470-58247-3.
58. Agresti, A. *Foundations of Linear and Generalized Linear Models*, 2nd ed.; Wiley: New York, NY, USA, 2015; ISBN 978-1-118-73003-4.
59. *Act No. 5/2004 Coll. on Employment Services and on Amending Certain Laws*; Ministry of Labour, Social Affairs and Family, Slovakia: Bratislava, Slovakia, 2004; Volume 2004.
60. Brownlee, J. *Classification Accuracy Is Not Enough: More Performance Measures You Can Use*; Machine Learning Mastery: San Juan, PR, USA, 2014.
61. Kroft, K.; Lange, F.; Notowidigdo, M.J. Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment. *Q. J. Econ.* **2013**, *128*, 1123–1167. [CrossRef]
62. Babos, P.; Lubyova, M. Effect of Labour Code Reform on Unemployment Duration in the Course of Crisis: Evidence from Slovakia. *Ekon. Cas.* **2016**, *64*, 218–237.
63. Lachiheb, A.B.A. Intermediation and Decision Support System for the Management of Unemployment: The Simulator of Duration. In Proceedings of the Digital Economy. Emerging Technologies and Business Innovation, Sidi Bou Said, Tunisia, 4–6 May 2017; Jallouli, R., Zaiane, O.R., Bach Tobji, M.A., Srarfi Tabbane, R., Nijholt, A., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 105–115.
64. Marksoo, U.; Tammaru, T. Long-Term Unemployment in Economic Boom and Bust: The Case of Estonia. *Trames* **2011**, *15*, 215. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.