ATOM – A FLEXIBLE MULTI-METHOD MACHINE LEARNING FRAMEWORK FOR PREDICTING OCCUPATIONAL SUCCESS

Bence Gergely
ELTE Doctoral School of Psychology
Károli University of the Reformed Church in Hungary
gergely.bence98@outlook.com

Szabolcs Takács
Károli University of the Reformed Church in Hungary
takacs.szabolcs.dr@gmail.com

Summary

Background and Aims: Presenting the statistical fundamentals of ATOM and its concurrent algorithms, with particular respect to demonstrate the flexibility of the decision-making module.
Methods: Simulating different classification problems using the Scikit Learn machine learning program package. During these simulations, the sample size, the number of variables, the number of groups, the proportion of incorrect classifications, and the distance between the groups were systematically changed.
Results: Based on 180 datasets, the Multilayer Perceptron performed the best in about 52% of the cases, and the Support Vector Classifier came in second place. It was found that every method proved to be better than any other in at least one case, which means that if we are dealing with a company or job where the given problem arises, these procedures provide a more accurate result. In addition, profound differences between different parameters of the same procedure were observed.
Discussion: Considering that the job selection aims to filter the best candidates, the accuracy of all procedures increases and, in general, it was shown that ATOM's algorithms indicate a performance much above the expected value of random categorization.
Keywords: recruitment automation, machine learning, psychological testing, multi-method approach

DOI: 10.17627/ALKPSZICH.2023.3.17

Introduction

Recruitment aims to provide the employer with appropriate human resources as cost-effectively as possible. Therefore, the selected employees must be able to perform the necessary tasks and have the cognitive and behavioural competencies required by the job (Hmoud & Varallyai, 2019). An effective selection process consists of several interrelated sub-processes but generally starts with defining the necessary tasks and abilities for the job, then continues with searching for and assessing the candidates, and ends with contracting the new employee (Ployhart, 2006). In the outlined process, human activity is essential since the selection cannot be, or is difficult to generalise. For the same reason, cognitive biases and heuristics decision-making are deeply rooted in selection (Whysall, 2018, Soleimani et al., 2022). For this reason, companies increasingly use recruitment software and attempt to partially or fully automate recruitment (Hmoud & Varallyai, 2019; Soleimani et al., 2022; Gonzalez et al., 2022; Liem et al., 2018).

Explanatory and predictive models

The primary goal of psychological science is to understand human behaviour (Yarkoni & Westfal, 2017). So, psychology primarily wants to explain phenomena with the simplest and most parsimonious models possible while placing less emphasis on prediction. So, in the vast majority of cases, psychology acts based on Occam's razor in model and theory formulation, i.e., it uses the most straightforward model with good explanatory power. The consequence is that the results can be only generalised within a closed theoretical framework and often have negligible predictive power (Robinaugh et al., 2021). In contrast, machine learning methods (especially deep neural networks) aim to maximise the prediction accuracy of the models. At the same time, mostly they do not provide an understandable explanation for how the phenomenon works (Yarkoni & Westfal, 2017). In that case, although it will provide a precise prediction for the given phenomenon, we will not (necessarily) know which variables and to what extent played a role in the outcome. Applied psychology often works with complex systems; therefore, the explanation of the processes is usually not the goal, mainly due to the scarcity of time and resources. Instead, the focus is on decision-making. Machine learning methods gained popularity in psychology, aiming to help professionals make decisions, such as in clinical diagnostics (Dwyer et al., 2018; Coutanche & Hallion, 2019). In the analysis of psychological experiments (Koul et al., 2018), academic success (Halde et al., 2016), and labour success (Liem et al., 2018).

The question is, do we want to understand the role of the factors involved during recruitment, or do we only want to provide a prediction? Suppose we only keep the explanation in mind. In that case, our selection process will probably be inflexible and not generalisable to other jobs, but we will earn a good understanding of the job's requirements. On the other hand, if we keep the prediction in mind and select the examined variables well in our model, our prediction will be accurate. However, if the variables are not appropriate, we will be unable to correct the prediction's inaccuracy.

In an ideal recruitment framework, one can optimise both aspects simultaneously:

giving a good prediction and indicating which variables play a role in the prediction come hand in hand (Kárász & Takács, this special issue).

Unique properties of recruitment data

The data arising from recruitment can be classified as 'soft' data. Its variability is much greater than data from physical measurement tools (Tannahil, 2007). Many times, this measurement error masks the otherwise complex data generation process. Due to the uncertainty of the variables, even complicated processes can appear linear (Yarkoni & Westfal, 2017). In order to reduce this uncertainty, work simulators and other instruments closer to physical measuring devices are often used in the field of work psychology (e.g., ErgoScope [Izsó, Berényi, & Takács, this special issue]).

The fact that it is difficult to access a large amount of data under given working conditions also contributes to the bias. Filling out long questionnaires can take a given employee out of production for several hours – however, it is difficult to make good predictions from a small amount of data (Yarkoni & Westfal, 2017). If going through the test battery takes a long time, missing data and systematic distortion of the test result occur more often (Nagybányai Nagy, 2013). That is why it is crucial to only ask for data that is needed – but we can only determine optimal test battery from preliminary measurements (Kárász & Takács, this special issue).

In addition, the quality of the data can also be questionable. There are often no established criteria for evaluating the performance of employees (Maji and Bera, 2020; van Esch et al., 2019). In many jobs, it is impossible to use objective performance indicators, and we can only obtain performance measures based on the subjective evaluation of HR professionals (Kárász & Takács, this special issue). Often, the selected psychological scales do not have predictive power, even in the case of high-reliability performance evaluations. The use of measurement tools is often limited to the kind of psychological tests the job has access to and whether they evaluate the effectiveness of the tests in the given recruitment process (Izsó, Berényi, & Takács, this special issue). The strength of the predictions largely depends on the quality of the input data. Analyses with low-quality data can raise serious validity problems – but these can also be handled to a significant extent by using different, more robust statistical procedures (Gergely & Vargha, 2021). It is often difficult to determine the quality of the data, but the multi-method approach adopted during the replication crisis can help a lot in drawing accurate conclusions. The essence of the multi-method methodology is that a given number of adequate statistical procedures are performed on a statistical question, then the obtained results are aggregated, thus making a more robust decision.

At the same time, the disadvantage of systems using more robust or complex methods is that the output data are difficult to interpret by professionals. Interpretability can be helped by providing the minimum information necessary for decision-making. For example (Izsó, Berényi, & Takács, this special issue) found that the feedback on the order of applicants and their classification into only two discrete acceptance categories can be sufficient for making a decision.

Decision-making module of ATOM framework

ATOM is a modular web-based framework that includes the compilation of the test battery, the organisation of recruitment campaigns, the analysis of the results, and the provision of automatic psychological feedback (Kárász & Takács, this special issue). Due to this structure, the goal of ATOM is to reduce the need for human resources in recruitment campaigns, thereby becoming a cheaper and more convenient alternative to traditional testing.

Training and test data requirements

ATOM's decision-making module is a flexible machine-learning framework combining several statistical methods. In each case, the input data consist of subscales of psychological and performance tests that have been validated and have high reliability. The subscale score is given by the sum of the items weighted to the subscale, which is standardised before the analyses (with a mean of 0 and a standard deviation of 1). The purpose of standardisation is to make the different subscales comparable, which is often a prerequisite for applied machine learning algorithms (Kárász & Takács, this special issue).

We need two types of input data to use the decision-making module: a training and a testing data file. In this case, the testing dataset represents the questionnaire results of the individuals applying for the given position, while the training dataset can be obtained from two sources. In the training data, we need information about whether an individual was proven to be a suitable candidate for a given job.

If the recruiting company has many employees, we can obtain the training data from the questionnaires filled out by these employees. Then these results must be labelled. Labelling means that the employees participating in the testing are classified into one of the predefined discrete groups (i.e., suitable, conditionally suitable, or not suitable for the position). These discrete groups can be created based on more objective performance measures (e.g., how many partners a sales employee contracts within a year), or the subjective evaluation of specialists can also provide the labelling. Expert evaluation is often fraught with cognitive biases, so to create optimal labelling, we need to ask for the opinions of several independent experts (Hallgren, 2012). Of course, the phenomenon of 'garbage in, garbage out' arises here, i.e., if the algorithms are trained with low-quality data, then the result (classification) will also be of poor quality. It is important to note that the decision-making module is structured in such a way that we can indicate the quality of the classification and the importance of the psychological and performance variables used, thus improving the efficiency of labelling and testing in the future.

If the recruiting company has few employees or there is no time for testing and labelling employees, then the quantification of expert opinions is a possible direction. Hmoud and Varallyai (2019) emphasise that the first step of the recruitment process is analysing the given job and assessing the necessary competencies. Hence, HR experts and work psychologists have a professional profile and optimally choose measurement instruments for this professional profile. ATOM's decision module can quantify this professional profile based on the measurement tools. First, the experts indicate which variables

are important, moderately important, or not important for the given job and also define the results required to be classified in the suitable candidate category. After that, we create a mixture of multidimensional normal distributions, which is parameterised based on the given expert values, whereas the non-determinable parameters (e.g., covariance) are fixed based on several different models (Gergely & Vargha, 2021). Labelling is defined here by belonging to a given component of the mixture distribution. The resulting artificial datasets reflect the expert opinion, but at the same time, they also include the uncertainty of the expert opinion.

It is important to note that the two student data file types cannot only operate independently of each other. For example, it may be the case that the company has few employees, but we take the tests with them, but to have a sufficient number of items, we also take the expert opinion into account.

Concurrent algorithms, hyperparameters, and cross-validation

If we have surveyed the employees or created the learning datasets, the next step is to fit the selected algorithms to the data. During the data analysis phase, the algorithms must predict the labels defined in the learning dataset, and the quality of the algorithm is determined by the accuracy of this prediction. The main idea behind ATOM's decision-making module is the use of concurrent algorithms, i.e., in contrast to the general practice (which specifies a model for the given use), several machine learning algorithms run in parallel, and the goal is to select the best solution for the given situation. The main advantage of competing algorithms is that they can adapt to the diversity of workplace selection, training data of varying size and quality, expert evaluation, and the specific characteristics of the job and latent data generation processes.

In order to optimally use and evaluate multiple algorithms together, three steps are required: hyperparameter setting, cross-validation, and measurement of the prediction accuracy.

Hyperparameters are the values that influence how a given algorithm works. Different algorithms can have different hyperparameters, and it is usually impossible to determine a combination of values that gives the best result in every case. In order to make it possible to measure which setting is the most optimal, we defined a hyperparameter space for each algorithm, with which we can determine which hyperparameter setting is the most suitable for the given problem by testing the algorithm with all possible hyperparameter combinations.

Some of the algorithms are not flexible. Logistic regression, being a generalised linear model, can fit one kind of function (a sigmoid function), while neural networks with different parameterisations can use many different non-linear functions. To take advantage of the strengths of the different algorithms, we use the method used for hyperparameter setting in this case as well. We create a model and hyperparameter list, the combination of which we fit the data and measure their effectiveness.

Machine learning algorithms learn based on how accurately they can predict the training dataset's labels. By increasing the flexibility of the procedure, we increase the possibility of overfitting. Overfitting means that the algorithm only learns the data, i.e., it will not

be able to reveal general patterns so that it will provide a suboptimal prediction in the case of previously unseen data. To minimise this possibility, we performed cross-validation on the entire model and hyperparameter space. The essence of cross-validation is to randomly divide the learning dataset into n equal parts and then create all possible (i.e., n pieces) partial learning datasets. The partial learning sets consist of n – 1 equal part, and the quality of the algorithm is tested only on the remaining one data part. This way, we test the algorithm's effectiveness on data that it has never seen before. We perform this process on all (n pieces) of the learning data set and then average the accuracy of the prediction, thus obtaining an estimate of how well the given algorithm performs on data it has not yet seen.

So far, we have not precisely defined what we mean by the efficiency of the algorithm and the quality of the prediction. There are several measures for this, depending on what we want to maximise/minimise in the given application. In this study, for the sake of simplicity, we used the percentage of correctly classified cases as an efficiency indicator. The percentage of correctly classified cases measures the percentage of predicted labels that match the actual labelling. In real selection situations, it makes sense to use several efficiency indicators, as the goal is usually not to categorise all applicants accurately but to filter out the best applicants. They will be forwarded to the interview process. In this case, a good efficiency indicator can be the percentage of correctly classified cases or the rate of false positives in the suitable candidate category. In summary, during the analysis phase, we select the algorithm-hyperparameter combination that receives the best score in the cross-validation procedure based on our determined efficiency measure.

Selected algorithms

During the construction of the decision-making module, the Python programming language was used in combination with the open-source Scikit-Learn program package (Python, 2021; Pedregosa et al., 2011).

Since we defined our dependent variable as a discrete category, we chose supervised learning algorithms that can solve classification problems. The complexity of the algorithms and the fact that the selected algorithms use different heuristics also played a role in the selection.

ATOM's decision-making module currently supports Logistic Regression (Wright, 1995), its regularised version (Cherkassky & Ma, 2003), the Support Vector Classifier algorithm family, Random Forest (Breiman, 2001), Adaboost (Freund & Schapire, 1997) and Multilayer Perceptron (Collobert et al., 2004).

Both the advantage and disadvantage of Logistic Regression lie in its simplicity: it is a generalised linear model capable of solving classification problems and requires few parameters for its operation. The Support Vector Classifier is a family of algorithms effective for multi-dimensional problems, even when the number of variables is larger than the number of sample elements. In addition, it is flexible since the function used for decision-making can be influenced by using different kernels. The disadvantage is that the probability of overfitting increases in the case of many variables. In such cases, regularisation and cross-validation should be used. Random Forest and Adaboost are

ensemble methods that combine several simple prediction algorithms (typically decision trees). While Random Forest is an averaging method that builds decision trees independently and then aggregates their results, Adaboost makes sequential estimates, i.e., builds a more efficient one from several weaker classification algorithms. Finally, the Multilayer Perceptron belongs to the family of artificial neural networks (ANN), but its version used in ATOM has only one hidden layer. The advantage of this solution is flexibility, its disadvantage is that it needs to estimate the weight and bias of the edges, which depends on the width of the input, output, and the hidden layer.

Output data and model evaluation

The final step in the decision-making module is to provide the output data. The most basic output data is the predicted labelling, and how the algorithms categorised the applicants. In some cases, this may be sufficient to select the applicants who enter the interview process, but the disadvantage is that it does not indicate how uncertain the decision was. The uncertainty of the decision can be quantified with labelling probabilities. When calculating the labelling probability, we do not classify the applicants under a label but give the probability of belonging to each group. For example, let us take two applicants; both of them were classified in the suitable category, but when we examine the labelling probability, we see that one belongs to the successful group with a 90% probability, while the other only with 65%.

In addition, we need to use measures that provide information about the performance of the models. Since not all methods can test the significance of the variables or indicate their importance, we used the Shap-value method (Shapley & Snow, 1952; Bowen & Ungar, 2020), which estimates the contribution of each variable to the prediction.

The purpose of this study is to demonstrate the flexibility of ATOM's decision-making module using simulations. In the case of different types of data occurring in the selection, the advantage of using several methods together prevails. So, in the case of simulated datasets, there will be at least one time when the given algorithm family gives the most accurate estimate, and the accuracy of the estimates will be similar between the models.

Methods

After the literature presentation and ATOM's methodology, the question may arise: Why is it necessary to use several concurrent process algorithms? In the machine learning literature, researchers traditionally present one procedure and compare it with algorithms created for a similar purpose or application. In this research, we want to show that using several simpler procedures (with fewer parameters) can achieve the robustness necessary to use data from psychological testing for recruitment.

The system is analysed using a simulation study. We created different classification problems during the simulation using the Scikit Learn machine learning program package (Pedregosa et al., 2011; Python, 2021). When creating the classification problems, we changed the size of the sample, the number of variables, the number of groups, the proportion of incorrect classifications, and the distance between the groups.

The sample size was 50, 100, 200, 500 and 1000, respectively, meaning that the total sample size for the training dataset was one of the values above. We considered that the sample size in psychology is often small and rarely exceeds 1,000 people. In addition, the sample size also reflects the number of individuals who can be tested on the Hungarian labour market; usually, medium-sized enterprises have around 50 employees, and in the case of large companies, it is not uncommon for a workforce of over 1,000 people (KSH, 2018).

The number of variables was divided into two categories: explanatory and redundant. Explanatory variables are those that can significantly predict which group the test person belongs to, while redundant variables are those that have no predictive power. The number of generated explanatory variables was 5, 10 and 20, respectively, for which we also created 10 redundant variables in each case. Redundant variables were considered important because it is common in workplace selection that some performance indicators do not have direct predictive power for the given job and are often used only because they are available or included in the test battery used by the company. An important question, in this case, is whether our automated procedures can filter these redundancies, thereby providing information about which variables should be used in the future.

The number of groups, i.e., the defined classes, was 2, 3 and 4. Here, we found that in practice, the inaccuracy of the grading increases as the number of categories increases in most companies. This is because the 2-point scale usually carries the essential information (suitable, not suitable candidate), and the 3-point scale (suitable, conditionally suitable, not suitable).

Table 1. Different parameters of the simulation setups

Parameters	**Values**				
Sample size	50	100	200	500	1000
No. variables	5	10	20		
Redundant variables	10				
No. groups	2	3	4		
Proportion of incorrect classifications	0.01	0.1			
Distance between groups	1	0.75			
Total	**180 classification problems**				

Source: created by the authors based on simulation details

We used the so-called incorrect classification ratio (0.01, 0.1), which means that 1% and 10% of the cases are already included incorrectly in the training dataset. Partly due to the inaccuracy of the suitability scale mentioned in the previous paragraph and partly due to the heuristic nature of human classification, we used these incorrect classification rates since it is assumed there are also false groupings in real datasets. This allows us to test the extent to which the learning algorithms can correct these evaluation biases.

The distance of the groups was set to 1 and 0.75, which means how ‘separated’

the clusters are from each other. A larger value means more separation, which results in an easier classification problem, while a smaller value means less separation and a more difficult classification problem. In a system where there are significant differences between suitable and not suitable candidates (1 standard deviation), we can expect significantly better results than in a case where the difference between them is smaller (0.75 standard deviations). Here, this should be understood as the number of standard deviation differences between the mean values when creating the mixture distributions.

We simulated a classification dataset with all possible combinations of these parameters, resulting in 180 different problems. Then, we ran the algorithms of ATOM framework with different parameterisations on each data file.

The effectiveness of the different algorithms and their different parameterisations was measured by the average accuracy of the classification (number of correct classifications / all cases). In this study, we used the first version of the ATOM, which only included accuracy as an outcome measure. For each algorithm, we present the number of cases when the given method provided the best accuracy. Moreover, we report the rank means over all 180 simulations.

To test and visualise the performance differences between algorithms and their different parameterisations, we perform a Kruskal-Wallis test with Bonferroni corrected pairwise comparisons and present the accuracy's median and the median absolute deviance.

Results

First, we consider the runtime of the simulations. The total runtime of the simulations study was approximately 11 minutes.[1] For the smaller datasets (50, 100) the grid search algorithm took only a few seconds (1-5s), none of the larger datasets took longer than 30 seconds to finish. Support Vector Classifier was the slowest to fit, albeit having the most parameters to sweep through with the grid search algorithm. Overall, we think that the speed of the algorithm is more than adequate for its use cases.

In the first step, we present which algorithm provided the most accurate prediction across all 180 datasets. In 51.1% of the cases, the Multilayer Perceptron, i.e., the neural network with one hidden layer, provided the best prediction, and the Support Vector Classifier came in second place. It is important to note here that the different parameterisations were not considered, the ratios here show how many times the given procedure provided the best prediction regardless of the different settings.

[1] All the simulations were running on a 2022 Apple Macbook Pro with an M1 Pro chip. The used packages were available for arm-type systems.

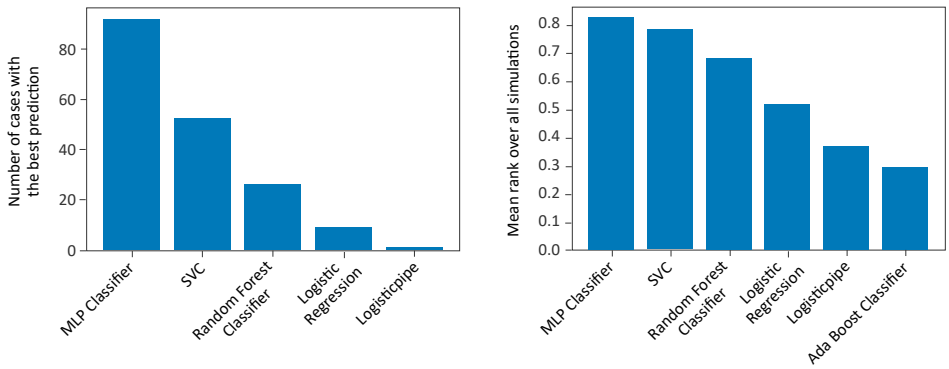

Figure 1. Number of cases when each algorithm provided the best prediction and the mean rank of the algorithms over all simulation
Source: the results were calculated and visualised using Python

However, the simulation aimed to show cases where it is unclear which procedure to choose. There were 10 classification problems each where Logistic Regression and the Logistic Pipe gave the most accurate prediction.

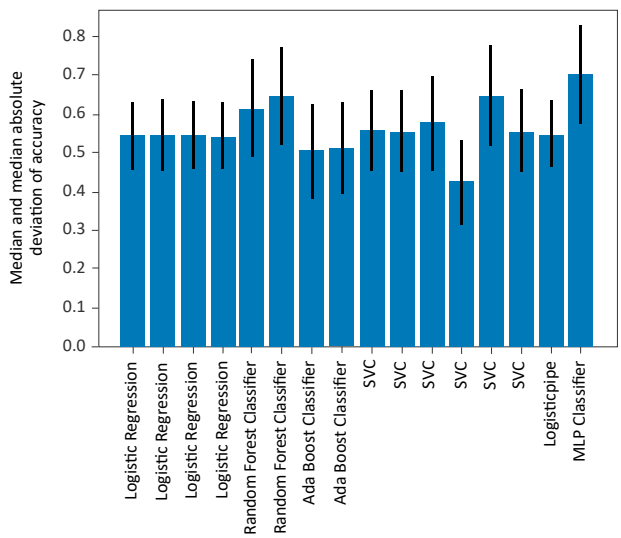

Figure 2. Median and median absolute deviation of accuracy for all algorithms and their parameterisations over all simulations
Source: the results were calculated and visualised using Python

That is, all methods except Adaboost proved to be better than any other in at least one case. This means that if we deal with a company or a job where the given problem arises, these procedures provide a more accurate result. However, the actual data cannot be analysed based on the simulation aspects since we usually have no information which procedure will be the most suitable before the analysis.

At the same time, we also examined the median accuracy of the procedures on all datasets and their median absolute deviations. There can be big differences even between different parameters of the same procedure. Based on the Kruskall-Wallis test [$H(5) = 111{,}656$; $p < 0.001$], and the pairwise comparisons, there are significant differences in the performance of the algorithms[2], and ultimately the algorithms can be ranked as Multilayer Perceptron, SVC, Random Forrest, Logistic Regression, Logistic Pipe and Adaboost, respectively.

Most importantly, all average results are typically above the 0.5 bands. This means that even in the case of a 2-valued prediction (success/failure), the prediction procedures have better results than completely arbitrary decision-making.

However, it is rarely important to classify each applicant accurately on the employer's side. It is much more important how well the given algorithm can guess the top 10% of applicants (the best 5–10–20 applicants), as these candidates will typically be the ones who will participate in the interview process.

Thus, we also looked at the median and median absolute deviations of the percentage of correctly classified cases for the top 10% of employees. In this case, the best method was the neural network: with a mean percentage of correctly classified cases of 70% and a standard deviation of 28%. So, if we are only interested in who the experts are, we can show an acceptable accuracy (in the case of 2 categories, we can show a rate of well over 50%).

Table 2. Post hoc comparison with Bonferroni correction

Post hoc comparisons - Accuracy							
		Mean Difference	SE	t	Cohen's d	p_{tukey}	p_{bonf}
AdaBoost-Classifier	Logistic-Regression	-0.046	0.011	-4.245	-0.274	< .001	< .001
	Logisticpipe	-0.047	0.015	-3.095	-0.283	0.024	0.030
	MLPClassifier	-0.141	0.015	-9.255	-0.845	< .001	< .001
	RandomForest-Classifier	-0.088	0.012	-7.074	-0.527	< .001	< .001
	SVC	-0.035	0.010	-3.477	-0.212	0.007	0.008

[2] Based on the Shapiro-Wilk test, the accuracies were likely non-normal in all cases.

Post hoc comparisons - Accuracy							
		Mean Difference	SE	t	Cohen's d	p_{tukey}	p_{bonf}
Logistic-Regression	Logistic-pipe	-0.001	0.014	-0.103	-0.009	1.000	1.000
	MLPClassifier	-0.096	0.014	-6.850	-0.571	< .001	< .001
	RandomForest-Classifier	-0.042	0.011	-3.924	-0.253	0.001	0.001
	SVC	0.010	0.008	1.297	0.062	0.787	1.000
Logistic-pipe	MLPClassifier	-0.094	0.018	-5.334	-0.562	< .001	< .001
	RandomForest-Classifier	-0.041	0.015	-2.681	-0.245	0.079	0.111
	SVC	0.012	0.013	0.881	0.071	0.951	1.000
MLP-Classifier	RandomForest-Classifier	0.053	0.015	3.479	0.318	0.007	0.008
	SVC	0.106	0.013	7.865	0.633	< .001	< .001
RandomForest-Classifier	SVC	0.053	0.010	5.187	0.316	< .001	< .001

Note: P-value adjusted for comparing a family of 6

Source: the results were calculated using JASP (Love et al., 2019)

Discussion

In this study, we presented the decision-making module of the ATOM framework and the advantage of the competitive algorithms method with the help of a simulation study.

ATOM's decision-making module was designed to answer the questions outlined in the introduction, namely the uncertainty coming from psychological assessment in recruitment scenarios. This uncertainty is often due to the small amount of data available for a given position. In case of a small sample size ATOM can quantify the expert evaluation of the different suitability categories and create a mixed dataset of actual and simulated candidates. Since companies rarely provide objective labelling of their employers, ATOM supports the good practice that HR experts independently assess the suitability of the employers. In the expert evaluation process, it is worth expecting high interrater reliability before starting the analysis. The goal is therefore not to exclude HR professionals from recruitment – but to best allocate their capacity and make the most optimal use of their expertise in the pre-screening and interview phase. In the case of longer-term cooperation, the amount of recruitment data is increasing over time, thus the prediction will be gradually better, but the quantification of expert opinion can reduce the cold start period, where our predictions are less than adequate. If a company is using ATOM for a longer period, it will gather data about not only the suitable candidates, but also the candidates who have not met the expectations. This way, the training sample becomes more representative of

the suitability categories; therefore, the algorithms will be more accurate overall.

It is important to note that the goal is not to accurately predict all categories of candidates. The goal is, instead, to identify who are the most suitable and likely to succeed, from whom the company can select the best candidates for the interview process. This process is facilitated by ATOM's decision-making module by freely changing the efficiency measures, thereby tailoring the analysis to the expectations of the job.

We selected algorithms that are not based on the same mathematical background; they require different assumptions and have varying robustness. That is, while the Support Vector Classifier is sensitive to the kernel type, it achieves good results in cases where the number of variables used is high. AdaBoost is not sensitive but tends to overfit in the case of many variables. At the same time, the power of the decision-making module is manifested in the fact that we do not have to take these assumptions into account, since these algorithms compete with each other on the training dataset, with different parameterisations and automatic model selection.

To account for the uncertainty of the outcomes, instead of just presenting the predicted suitability, we also report the probabilities of belonging to each category. In this way, employers can create their own rankings: filtering the least likely succeeding candidates or selecting the most potent ones (Izsó, Berényi, & Takács, this special issue).

Based on the simulation, we can say, that in the case of our developed system, the selected algorithms create a flexible framework. Moreover, all algorithms, except the Adaboost provided the best prediction in at least one case. Nevertheless, it was expected that the neural network would produce the best results due to the algorithm's robustness (Collobert & Bengio, 2004).

Concerning the accuracy of the prediction, we did not experience any substantial differences between the different methods and their different parameterisations. The average performance was above 50% respectively. Note that the expected value of a completely random selection is 35%. So, each algorithm results in a much more accurate categorisation on average. If we consider that the job selection aims to filter the best candidates, the accuracy of all procedures increases and, in general, our algorithms show a performance well above the expected value of random categorisation (rate of 35%).

The current algorithm's limitation is that it selects a single model in each case and does not account for the strength of different models. A model selection resembling the Bayesian model averaging would be more suitable than choosing the most accurate model. Furthermore, the algorithm's flexibility needs to be further assessed with different types of data and jobs.

The limitation of the simulation study is that the simulated datasets all came from a mixture of normal distributions, with equal distances between the centroids of the components.

Összefoglaló

ATOM – egy rugalmas, több módszert alkalmazó gépi tanulási keretrendszer a munkahelyi beválás előrejelzésére

Háttér és célkitűzések: Jelen kutatás bemutatja az ATOM szoftvert és annak statisztikai megfontolásait, különös tekintettel a döntéshozatali modul rugalmasságának demonstrálására. *Módszer*: Scikit Learn segítségével különböző osztályozási problémákat szimuláltunk. A szimulációk során szisztematikusan változtattuk a minta méretét, a változók számát, a csoportok számát, a hibás osztályozások arányát és a csoportok közötti távolságot. *Eredmények*: A 180 szimulált adatállomány alapján a Multilayer Perceptron az esetek mintegy 52%-ában a legjobban teljesített, a második helyen pedig a Support Vector Classifier végzett. Megállapítottuk, hogy minden módszer legalább egy esetben jobbnak bizonyult a többinél, ami azt jelenti, hogy ha olyan céggel vagy munkakörrel foglalkozunk, ahol az adott probléma felmerül, akkor ezek az eljárások pontosabb eredményt adnak. Ezenkívül lényeges különbségeket figyeltünk meg ugyanazon eljárás különböző paraméterezései között. *Következtetések*: Tekintettel arra, hogy a kiválasztás célja a legjobb jelöltek kiszűrése, az összes eljárás pontossága növekszik, ha csak a legegyertelműbben kategorizálhatókat keressük. Általánosságban megmutatkozott, hogy az ATOM algoritmusai a véletlenszerű kategorizálás várható értékét jóval meghaladó teljesítményt jeleznek.
Kulcsszavak: munkaerő-kiválasztás automatizációja, gépi tanulás, pszichológiai tesztelés, konkurens algoritmusok alkalmazása

References of this Special Issue

Izsó, l., Berényi, B., & Takács, Sz. (2023). Illustrating real-life ATOM application case studies. *Alkalmazott Pszichológia*, *25*(3), 93–114.

References

Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. *arXiv*. https://arxiv.org/abs/2006.07155

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

Cherkassky, V., & Ma, Y. (2003). Comparison of model selection for regression. *Neural Computation*, *15*(7), 1691–1714. https://doi.org/10.1162/089976603321891864

Collobert, R., & Bengio, S. (2004). Links between perceptrons, MLPs and SVMs. *Proceedings of the Twenty-first International Conference on Machine Learning*, 23. https://doi.org/10.1145/1015330.1015415

Coutanche, M. N., & Hallion, L. S. (2020). Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright & M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology.* Cambridge University Press. https://doi.org/10.1017/9781316995808.041

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Gergely, B., & Vargha, A. (2021). How to Use Model-Based Cluster Analysis Efficiently in Person-Oriented Research. *Journal for Person-Oriented Research*, *7*(1), 22–35. https://doi.org/10.17505/jpor.2021.23449

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., & Martin, N. R. (2022). Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior, 130*(May), 107179. https://doi.org/10.1016/j.chb.2022.107179

Halde, R. R., Deshpande, A., & Mahajan, A. (2016). Psychology assisted prediction of academic performance using machine learning. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 431–435. https://doi.org/10.1109/RTEICT.2016.7807857

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–43. https://doi.org/10.20982/tqmp.08.1.p023

Hmoud, B. I., & Várallyai, L. (2020). Artificial intelligence in human resources information systems: Investigating its trust and adoption determinants. *International Journal of Engineering and Management Sciences*, *5*(1), 749–765. https://doi.org/10.21791/IJEMS.2020.1.65

KSH (Központi Statisztikai Hivatal) (2018). Munkaerőpiaci helyzetkép, 2014–2018. http://www.ksh.hu/docs/hun/xftp/idoszaki/munkerohelyz/munkerohelyz17.pdf

Koul, A., Becchio, C., & Cavallo, A. (2018). PredPsych: A toolbox for predictive machine learning-based approach in experimental psychology research. *Behavior Research Methods*, *50*(4), 1657–1672. https://doi.org/10.3758/s13428-017-0987-2

Liem, C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1–17. https://doi.org/10.18637/jss.v088.i02

Nagybányai Nagy, O. (2013). *The Effect of Response Style Characteristics on the Measuring Efficiency of Self-administered Testing Methods* [Doctoral dissertation]. Eötvös Loránd University.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*(6), 868-897.

Python, W. (2021). Python. *Python Releases for Windows*, *24*.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Shapley, L. S., & Snow, R. N. (1952). Basic solutions of discrete games. *Contributions to the Theory of Games*, *1*, 27–35. https://doi.org/10.1515/9781400881727-004

Soleimani, M., Intezari, A., & Pauleen, D. J. (2022). Mitigating cognitive biases in developing AI-assisted recruitment systems: A knowledge-sharing approach. *International Journal of Knowledge Management*, *18*(1), 1–18. https://doi.org/10.4018/IJKM.290022

Tannahill, G. K. (2007). *A study of soft skills for IT workers in recruitment advertising.* Capella University

van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222. https://doi.org/10.1016/j.chb.2018.09.009

Whysall, Z. (2018). Cognitive biases in recruitment, selection, and promotion: The risk of subconscious discrimination. In V. Camen, & S. Nachmias (Eds.), *Hidden inequalities in the workplace: A guide to the current challenges, issues, and business solutions* (pp. 215–243). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-59686-0_9

Wright, R. (1995). Logistic regression. *Reading and Understanding Multivariate Statistics*, 217–244.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393