Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 103-107, 11, 2023

Araştırma Makalesi

IJANSER
https://alls-academy.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 103-107, 11, 2023

Research Article

A Survey of The State-Of-The-Art AutoML Tools and Their Comparisons

Ahmet Serhat Fidan [1], Murat Şimşek [2*] and Buğra Kağan Kayhan [3]

[1]*Computer Engineering, Yalova University, Turkey*
[2]*Artificial Intelligence Engineering, Ostim Technical University, Turkey*
[3]*Software Engineering, Ostim Technical University, Turkey*

[*]*murat.simsekstimteknik.edu.tr*

(Received: 03 December 2023, Accepted: 11 December 2023)

(2nd International Conference on Frontiers in Academic Research ICFAR 2023, December 4-5, 2023)

ATIF/REFERENCE: Fidan, A. S., Şimşek, M. & Kayhan, B. K. (2023). A Survey of The State-Of-The-Art AutoML Tools and Their Comparisons. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(11), 103-107.

Abstract – Machine learning is used effectively in many areas today and its usage area is increasing day by day. In addition, processes based on machine learning are also developing in a technology-oriented manner, and users are gaining new perspectives on solving current problems. While machine learning makes predictions about stocks in the financial sector, it also plays an active role in early diagnosis of diseases in the healthcare sector. It is actively used in route calculation and defective product detection in the field of production and logistics, and in situations such as analysis of customer behavior and product recommendations in the shopping sector. AutoML can be defined as a process that aims to automate the machine learning process end-to-end. It enables the machine learning process to be accelerated by automating especially time-consuming tasks that work with the logic of repetition, and it allows people who work in this field to create more efficient and productive models. In addition, AutoML helps users who are not experts in this field in the stages of machine learning model development, data management, analysis and evaluation of their own data, by providing various conveniences to users in model training and subsequent stages. In this article, after discussing what AutoML is, AutoML processes and areas of use, information about various AutoML platforms have been given, the differences between widely used AutoML platforms will be evaluated, and their advantages and disadvantages compared to each other will be included.

Keywords – Machine Learning, Automated Machine Learning, AutoML, Hyperparameter Optimization, Model Selection

I. INTRODUCTION

Machine learning (ML) has recently pushed its way into our daily lives. ML can assist in suggesting to the active user what to read, Which movies to watch, etc. ML can be used to predict certain things and guide users. However, this widespread application has also demonstrated that human professionals must have extensive expertise and effort in order to use it well [1]. AutoML is a platform that automates certain stages in the ML process. Automating these processes in machine learning not only makes these platforms easy for users to use, but also contributes to a more consistent and diverse presentation of results [2]. With this, AutoML is a platform that makes machine learning more usable and enables people and institutions who are not experts in this field or do not have sufficient knowledge to use ML effectively [3].

Automated Machine Learning (AutoML) has become a rapidly developing field in recent years

[4]. Individuals interested in data science and machine learning can automate processes such as data preprocessing, model selection, hyperparameter optimization, and model interpretation using AutoML tools to save time and effort, especially on repetitive tasks. In this way, it is possible to obtain better results by making the machine learning process more efficient.

Figure 1 shows a standard Machine Learning Process which includes a data scientist to complete tasks. AutoML aims to minimize the need for the expertise of a Data Scientist.

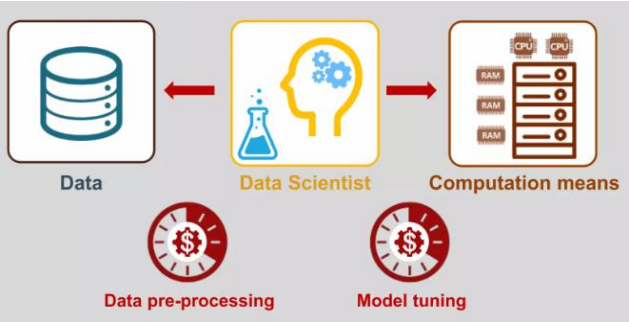

Fig. 1 Standard Machine Learning Process [5]

In this article, the current state of AutoML tools was examined and a comparison was made between different tools. In other words an end to end machine learning application necessity comparison is made. Using an iterative or linear pipeline configuration and viewing the ML process of any problem that contains data as an optimization problem has a substantial advantage and AutoML's objective is not to entirely replace people in analysis. In order to achieve a more efficient process, man and machine must work together [6].

AutoML tools include commercial tools such as H2O-AutoML, while open-source tools such as Auto-Keras, Auto-sklearn and TPOT are also available. While commercial tools generally have more detailed analysis and interpretation features, open-source tools focus more on model selection and hyperparameter optimization. In this article, general information about the current status and usage areas of AutoML tools was given. In the following sections, the advantages and disadvantages of different AutoML platforms were evaluated.

The following sections contain a detailed review of AutoML tools and an evaluation of their performance. This article aims to provide a general perspective on the current state of AutoML tools and to guide those who are dealing with this field or want to gain knowledge.

II. AUTOML PROCESS

AutoML processes consist of a series of steps used to automate repetitive tasks in machine learning projects. These steps include data preprocessing, model selection, hyperparameter optimization, and model interpretation.

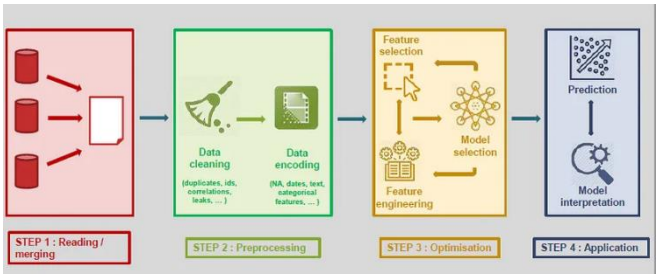

Fig. 2 AutoML Processes [5]

Data preprocessing is the first and most important step in the machine learning process. In this step, the available data set is cleaned, missing data is filled and unnecessary data is removed [7]. Additionally, the data set can be normalized or various transformations can be performed. Currently, this task cannot be performed effectively by any of the existing AutoML tools. Therefore, significant human intervention is required in this process. Especially during this process, existing AutoML tools need data type and schema detection, which are not widely supported.

Model selection is the next step in the AutoML process. In this step, the most appropriate ML algorithms to be used for the target job are selected. One of the challenges in choosing the best algorithm is managing the balance between accuracy and interpretability [8]. This selection is usually made using a performance metric. AutoML tools can automatically or manually select the best model based on user-specified performance metrics.

Hyperparameter optimization aims to find the best values of parameters that affect the performance of machine learning models. These parameters significantly affect the complexity, learning rate, and overall performance of the model. AutoML tools typically find the best

parameter values at this stage using automatic hyperparameter optimization algorithms.

Finally, the model interpretation and result analysis step constitutes the final stage of the AutoML process. In this step, the performance of the selected model is evaluated and the results are evaluated. AutoML tools can offer a variety of analysis methods to understand how the model makes decisions. These analysis methods can be used to identify important features of the model or to visualize how the model works.

If used effectively, the mentioned AutoML processes not only save time and effort in machine learning projects, but also make it easier to obtain more stable results. AutoML tools help users create machine learning models faster, more effectively, and more consistently. Therefore, understanding AutoML processes and choosing platforms that use them correctly and effectively gives the user a significant advantage in machine learning projects.

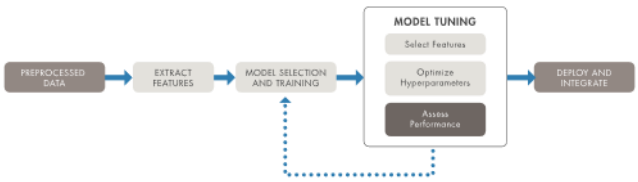

Fig. 3 Streamlining machine learning workflows with AutoML [9]

III. AUTOML TOOLS AND THEIR FEATURES

In this part of the article, the features of some AutoML tools will be mentioned with comments. Commonly used AutoML platforms include H2O AutoML, Auto-Keras, Auto-WEKA Auto-sklearn and TPOT [3,10].

A. *H2O AutoML*

H2O AutoML is a commercial and open source AutoML tool that can automatically detect the data schema and provide detailed result analysis [11]. It also has special functionality for processing time series data.

H2O AutoML automates all feature engineering and turns features into values that ML algorithms can easily process. Automates time-consuming data science tasks and creates pipelines with low latency. It makes comparisons to find the best model by iteration. On top of its ML Interpretability and fairness dashboards it includes automated model documentation and reason codes for each model prediction, providing what is needed to build a secure machine learning lifecycle. It also uses an AI wizard that explores your data, making recommendations based on your business needs, and giving instructions on appropriate machine learning techniques to use based on your data and use case needs [12].

B. *Auto-Keras*

Auto-Keras is an open source AutoML tool and is built on top of the Keras library. It is capable of automatic neural network research, especially for image and text data [13]. Using Auto-Keras, one can create a model that includes complex elements such as embeddings and reduction techniques that are normally less accessible to those still learning Deep Learning. At the same time, a neural architecture search algorithm finds the best architectures, such as the number of neurons in a layer, the number of layers, the layers to include. When a model is created with Auto-Keras, the model can be used as a regular TensorFlow/Keras model, it does a lot of pre-processing automatically, such as vectorizing or cleaning text data [14].

C. *Auto-WEKA*

Auto-WEKA works by implementing WEKA's classification and regression algorithms in other words it's built on WEKA [15]. WEKA is a widely used, open-source ML platform. Thanks to its user-friendly interface, it is especially popular with novice users. However, novice users still find it difficult to choose the best approach for their own data. Auto-WEKA is a system designed to assist such users by automatically adjusting WEKA's learning algorithms and hyperparameter settings to maximize performance using the Bayesian Optimization method. Auto-WEKA is tightly integrated with WEKA, making it as accessible to its end users as any other learning algorithm [16].

D. *Auto-Sklearn*

Auto-sklearn, is another open source AutoML tool. It is Python-based and built on Scikit-learn library and it uses the well-known Scikit-Learn machine learning package for data processing and ML algorithms [17]. Auto-sklearn automatically searches for the right learning algorithm for a new machine learning dataset and optimizes its hyperparameters. It also includes a Bayesian

Optimization search technique to quickly find the best pipeline for a given data set [18]. Auto-sklearn extends the idea of structuring a general ML framework with efficient global optimization introduced with Auto-WEKA. Auto-sklearn uses the same Bayesian Optimization method as Auto-WEKA, but since scikit-learn does not apply as many different ML techniques as WEKA, it includes a smaller model and hyperparameter space and uses meta-learning to identify similar datasets and use information collected in the past. It covers a total of 15 classification algorithms, 14 feature preprocessing algorithms and deals with data scaling, coding of categorical parameters and missing values [10].

E. TPOT

TPOT, short for Tree-based Pipeline Optimization Tool, is an open-source tool for performing AutoML in Python. TPOT uses a tree-based structure, hence the name, to represent a model pipeline for a predictive modeling problem, including data preparation and modeling algorithms and model hyperparameters [19]. It also uses genetic programming to efficiently discover the best-performing model pipeline for a given data set. Genetic programming is a methodical approach to getting computers to automatically solve problems [20]. Additionally, while requiring little to no user input or prior knowledge, TPOT may create machine learning pipelines that significantly outperform a basic machine learning analysis. And through incorporating Pareto optimization, which creates compact pipelines without compromising classification accuracy, TPOT has a tendency to generate overly complicated pipelines [21].

IV. RESULTS

Each AutoML tool has slight differences, but in general they all perform hyperparametric optimisation, select the appropriate ML algorithm for the data set, build the appropriate pipeline and help even a novice user to analyse data through ML. Especially H2O AutoML is one of the most user-friendly tools. Since this tool is a commercial AutoML tool, it aims to maximise the user experience. Although nothing can be said for certain in terms of the accuracy of the results, open source AutoML tools such as Auto-Keras, Auto-WEKA Auto-sklearn and TPOT are likely to give better results. Since user control is higher in open-source tools, it may allow the desired work to be done exactly, but the number of errors that may come with control tends to increase. In addition to these features, other feature comparisons of some of these tools can be seen in Table 1.

Table 1. Comparison of AutoML Tools

	Auto-Sklearn	**TPOT**	**Auto-Keras**	**Auto-WEKA**	**H2O AutoML**
Data Preprocessing	No	No	No	No	Yes
Automatically Detecting Data Types	No	No	No	No	Yes
Unsupervised Learning	No	No	No	No	No
Flexible Parameter Selection	No	Yes	Yes	Yes	No
Supervised Learning	Yes	Yes	Yes	Yes	Yes
Handling Imbalanced Feature Dataset	Yes	Yes	Yes	Yes	Yes
Ensemble Learning	Yes	Yes	No	Yes	Yes

Table 1 presents a comparative analysis of frequently utilized AutoML tools within the literature. However, as evident from the table, existing AutoML tools still exhibit numerous shortcomings. This table has primarily been created to illustrate the deficiencies inherent in AutoML tools, aiming to shed light on areas that require further attention for those engaged in AutoML research. Its purpose is to serve as an informative resource regarding the facets of AutoML that necessitate closer examination.

V. DISCUSSION

One important conclusion from the study was the realization that while AutoML unquestionably improved machine learning's usability, it was not eliminate the need for human expertise. This emphasizes the value of working together between AutoML tools and subject matter experts who can offer crucial context, problem framing, and interpretation of results. AutoML can be viewed as a powerful assistant to data scientists, helping them streamline repetitive tasks and optimize models, but it cannot replace the creativity and domain

knowledge that human professionals bring to the table.

The findings also point out the significance of data pre-processing as the initial step of the machine learning pipeline. It is pointed out that current AutoML tools lack the power to automate this phase effectively. As automating this phase can significantly reduce the need for experts and expertise future AutoML development should focus on improving data preprocessing automation. Additionally, finding the right balance between model accuracy and interpretability is still a point of discussion thus research into more interpretable, yet highly accurate models can be looked more into.

Overall, the foundings highlight the need for ongoing collaboration between AutoML tools and domain experts and identify important areas for improvement, such as data preprocessing automation and model interpretability. As AutoML continues to evolve, it has the power to spread machine learning and its accessibility to a wider audience by leveraging the experience of experts in the field.

REFERENCES

[1] Vaccaro L., Sansonetti G., Micarelli A. An empirical review of automated machine learning.

[2] Zöller MA., Huber MF. Benchmark and survey of automated machine learning frameworks. Journal of Artificial Intelligence Research 2021; 70: 409-472.

[3] Thornton C., Hutter F., Hoos HH., Leyton-Brown K. Auto-weka: combined selection and hyperparameter optimization of classification algorithms. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 847-855, 2013.

[4] Guyon, I., Sun-Hosoya L., Boullé M., Escalante HJ., Escalera S., Liu Z., Jajetic D., Ray B., Saeed M., Sebag M., Statnikov AR., Tu W-W., Viegas E. Analysis of the automl challenge series 2015-2018. NeurIPS Workshop Proceedings, pp. 177-219, 2019.

[5] (2023) Slideshare website. [Online]. Available: https://www.slideshare.net/AxeldeRomblay/automate-machine-learning-pipeline-using-mlbox

[6] Özdemir Ş., Örslü S. Makine öğrenmesinde yeni bir bakış açısı: otomatik makine öğrenmesi (automl). Journal of Information Systems and Management Research, 1 (1): 23-30, 2019.

[7] Jesmeen M., Hossen J, Sayeed S., Ho C. A survey on cleaning dirty data using machine learning paradigm for big data analytics. Indonesian Journal of Electrical Engineering and Computer Science, 10(3): 1234-1243, 2018.

[8] Operskalski JT., Barbey AK. Risk literacy in medical decision-making. Science, 352(6284): 413-414, 2016.

[9] (2023) Matworks website. [Online]. Available: https://www.mathworks.com/discovery/automl.html

[10] Feurer M., Klein A., Eggensperger K., Springenberg J., Blum M., Hutter F. Efficient and robust automated machine learning. Advances in neural information processing systems, 28: 2962–2970, 2015.

[11] Darren C. Practical machine learning with H2o: powerful, scalable techniques for deep learning and ai. O'Reilly Media, Inc. 2016.

[12] LeDell E., Poirier S. H2o automl: scalable automatic machine learning. 7th ICML Workshop on Automated Machine Learning 2020.

[13] Jin H., Song Q. Auto-keras: an efficient neural architecture search system. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1946–1956, 2019.

[14] Jin H., Chollet F., Song Q., Hu X. Autokeras: an automl library for deep learning. Journal of Machine Learning Research, 24(6): 1-6, 2023.

[15] Kotthoff L., Thornton C., Hutter F. User guide for auto-weka version 2.6. Dept. Comput. Sci., Univ. British Columbia, BETA Lab, Vancouver, BC, Canada, Tech, Rep 2: 1-15, 2017.

[16] Kotthoff L., Thornton C., Hoos HH. Auto-weka 2.0: automatic model selection and hyperparameter optimization in weka. Journal of Machine Learning Research, 17: 1-5, 2016.

[17] Pedregosa F., Varoquaux G., Gramfort A., Scikit-learn: machine learning in python. Journal of Machine Learning Research, 12: 2825-2830, 2011.

[18] Feurer M., Eggensperger K., Falkner S., Lindauer M., Hutter F. Auto-sklearn 2.0: hands-free automl via meta-learning. The Journal of Machine Learning Research, 23(1): 1-61, 2022.

[19] Olson RS., Urbanowicz RJ., Andrews PC, Lavender NA., Kidd LC., Moore JH., Automating biomedical data science through tree-based pipeline optimization. Applications of Evolutionary Computation: 19th European Conference, EvoApplications, pp.123-137, 2016.

[20] Koza JR., Poli R. Genetic programming. In: Burke, E.K., Kendall, G. (eds) Search Methodologies. Springer, Boston, MA, pp. 127-164, 2005.

[21] Olson RS., Bartley N., Urbanowicz RJ., Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. Proceedings of the Genetic and Evolutionary Computation Conference, pp. 485–492, 2016.