Production Time Prediction for Contract Manufacturing Industries Using Automated Machine Learning

Afonso Sousa[1,3][0000−0001−6336−6497], Luís Ferreira[1,3][0000−0002−4790−5128], Rui Ribeiro[1,3][0000−0001−8078−4148], João Xavier[2], André Pilastri[3][0000−0002−4380−3220], and Paulo Cortez[1][0000−0002−7991−2090]

[1] ALGORITMI R&D Centre, Department of Information Systems, University of Minho, 4804-533 Guimarães, Portugal
pcortez@dsi.uminho.pt
[2] Colep Packaging, 3730-423, Vale de Cambra, Portugal
joao.xavier@colep-pk.com
[3] EPMQ - IT CCG ZGDV Institute, 4804-533 Guimarães, Portugal
afonso.sousa@ccg.pt, luis.ferreira@ccg.pt, rui.ribeiro@ccg.pt, andre.pilastri@ccg.pt

Abstract. The estimation of production time is an essential part of the manufacturing domain, allowing companies to optimize their production plan and meet the dates required by the customers. In the last years, there have been several approaches that use Machine Learning (ML) to predict the time needed to finish production orders. In this paper, we use the CRISP-DM methodology and Automated Machine Learning (AutoML) to address production time prediction for a Portuguese contract manufacturing company that produces metal containers. We performed three CRISP-DM iterations using real data provided by the company related to production orders and production operations. We compared four open-source modern AutoML technologies to predict production time across the three iterations: AutoGluon, H2O AutoML, rminer, and TPOT. Overall, the best results were achieved in the third CRISP-DM iteration by the H2O AutoML tool, which obtained an average error of 3.03 days. The obtained results suggest that the inclusion of data about individual manufacturing operations is useful for improving production time for the entire production order.

Keywords: Contract Manufacturing · Automated Machine Learning · Regression.

1 Introduction

Under the current highly competitive market, manufacturing companies are constantly being challenged by growing demands for individualized products that present a higher quality and simultaneously a lower production cost. In particular, modern logistics methods, such as high on-time or short delivery and lead

times, are a way to differentiate the companies from their competitors [16]. In general, the manufacturing process consists of the transformation of raw materials into components or final products, through a sequence of operations [13]. Since a production plan can contain several production orders with different operations, the exact prediction of the production time for an order plays a determining role in improving the efficiency of the plan, allowing to meet the dates required by the customer [11].

This paper presents an implementation of the CRISP-DM methodology and Automated Machine Learning (AutoML) to predict production time in the manufacturing domain. We used real-world data collected from a Portuguese contract manufacturing company that produces metal containers, such as cans and aerosols. The prediction of production time is essential for the company, in order to optimize the production plan.

Our approach consisted of using several CRISP-DM iterations to provide predictions of the time needed to conclude a production order. We used four recent open-source AutoML tools for the CRISP-DM modeling phase, to automate algorithm selection and hyperparameter tuning: AutoGluon, H2O AutoML, rminer, and TPOT. The usage of AutoML allowed us to focus on the data preparation phase. We executed a total of three CRISP-DM iterations, in which we tested different data preprocessing and feature engineering techniques.

The paper is structured as follows. Section 2 presents the background work. Section 3 describes the CRISP-DM methodology and the AutoML tools. Section 4 details the three CRISP-DM iterations we performed. Then, Section 5 presents the overall results and discussion. Finally, Section 6 presents the main conclusions and future work directions.

2 Background

The lithography process of the company begins with cutting the coil. Given the diversity of containers that the company produces, there are numerous metal and cutting specifications to meet the product requirements. The next step is the varnishing process. In this process, varnish or enamel is applied to the metal sheet, to protect the can from the erosion caused by its contents. Next, the sheet goes through a printing process. This stage consists of applying ink to a certain area of the sheet, translating the color and illustration according to the client's request. Finally, the sheet goes through a secondary cutting process, transforming it into several bodies or strips.

Each step of the lithography factory process has a set of associated operations that influence production time. Through a predictive modeling approach, we can increase the accuracy of forecasts for the production times of an order, which is a key indicator for the development of a successful production plan [8]. Although there are research works on the use of Data Mining techniques for the forecasting of manufacturing order delivery times, within our knowledge the research concerning data-driven modeling of lithography process is scarce.

In [1], the authors estimated the weekly lead times of a producer of heavy electric motors by using Artificial Neural Networks (ANN), fuzzy regression, and a traditional regression. The obtained results showed that ANN is superior to both regression and fuzzy regression, obtaining lower Mean Absolute Percentage Error (MAPE) values. ANN were also used by [9]. The authors used technical specifications demanded by the customer to predict the flow time of a wide range of products, obtaining an accuracy superior to 90% when comparing the results of ANN and factorial design with check data. [8] compared a traditional Operations Research (OR) approach with ML approaches and concluded that ML provides more precise results than OR when predicting the production lead time of optical lenses for eyeglasses.

3 Materials and Methods

3.1 CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) is a framework that allows translating a business problem into a Data Mining project, providing a holistic view of the life cycle of a Data Mining project, beyond the application of ML models. It is divided into six different phases, that cover further general tasks [17]. First, in the business understanding phase it is necessary to understand what the goals are from a business point of view and to fit these requirements into a Data Mining problem. In the data understanding phase, it is essential to analyze the collected data in order to draw insights and identify possible problems. The data preparation phase processes all relevant data to generate the final dataset. In the modeling phase, various techniques and algorithms are applied to extract knowledge from the processed data. In the evaluation phase, the generated model is typically tested with real data, to ensure the quality of the model. Finally, the deployment phase can vary depending on the project requirements. It can be a final report or the implementation of a repeatable Data Mining process [2].

3.2 AutoML

The modeling phase of CRISP-DM is mainly related to the choice of a modeling technique, usually involving ML algorithms (e.g., Linear Regression, Decision Trees) and their hyperparameters. Typically, this is a very iterative phase, that might require trial and error approaches, domain knowledge, or heuristics [7]. Given the iterative nature of this step, this can become a time-consuming process. Automated Machine Learning (AutoML) helps reduce the time needed for the modeling phase of CRISP-DM by choosing the best algorithm and hyperparameters for a given dataset, without the need for human input [6].

In this work, we use AutoML to automate algorithm selection and hyperparameter tuning for the modeling phase of CRISP-DM. The usage of AutoML intends to reduce the time needed for the modeling phase and allowed us to focus on other key phases, such as data understanding and data preparation [15].

For our experiments, we used four open-source AutoML tools: AutoGluon, H2O AutoML, rminer, and TPOT. AutoGluon is the AutoML module for the Gluon framework [4]. For our experiments, we used the tabular prediction component of AutoGluon. H2O AutoML is the open-source AutoML module from the H2O framework [10]. Rminer is a package for the R language that facilitates the usage of ML algorithms and also includes flexible AutoML functions [3]. For rminer, we used the `"automl3"` template[4]. Lastly, TPOT is an AutoML library based on the Scikit-Learn framework that uses Genetic Programming [14]. Table 1 summarizes the characteristics of the four used AutoML technologies (**Framework**, **API** language, and used **Version**).

Table 1. Description of the adopted AutoML tools.

Tool	**Framework**	**API**	**Version**
AutoGluon	Gluon	Python	0.3.1
H2O AutoML	H2O	Java, Python, R	3.36.0.3
rminer	rminer	R	1.4.6
TPOT	Scikit-Learn	Python	0.11.7

4 CRISP-DM Methodology

4.1 First CRISP-DM Iteration

In the first iteration, we performed the first five phases of CRISP-DM (from Business Understanding to Evaluation). Our goal was to predict production time, in other words, the time needed to conclude a production order.

Business Understanding In this phase, the company identified the need to predict when a production order will be finished, given a defined starting date. We identified a regression task, in which we aim to predict the number of days needed to finish a production order after it starts.

Data Understanding For this phase, we analyzed the first data source provided by the company: Production Orders. This dataset is composed of 14,614 records, gathered from December 2018 to December 2020. Each record is related to a finished production order. Table 2 summarizes the attributes of the dataset.

[4] https://CRAN.R-project.org/package=rminer

Table 2. Description of the Production Orders dataset attributes.

Attribute	Description	Type	Levels	Example
Order_ID	ID of the production order	Categorical	14438	1000234997
Material	Material of the final product	Categorical	4490	51-49012
Order_quantity	Produced quantity of order	Numerical	-	423
Start_order	Starting date of production order	Date	394	28/03/2019
End_order	Ending date of production order	Date	341	06/01/2020

Data Preparation First, we created the target variable: the number of days needed to finish each production order, which we named Delta. This new column resulted from the subtraction of the columns End_order and Start_order (in total days since the company also works at the weekends). For the data preprocessing, we deleted records with missing values (e.g., records without a registered material or with missing dates). There were instances of records with duplicate Order_ID (around 175 records). In these cases, we deleted the duplicate records and only considered the first record with a distinct ID. Then, we removed the column Order_ID since it has no predictive value. Additionally, we removed the target column outliers. The company suggested that records with production time higher than 150 days should be considered outliers, so we removed any record with a Delta over 150. We noticed some records had negative Delta and we also remove these records.

Finally, we applied a standardization technique to the only numeric column (Order_quantity), which transformed the values into a new scale with a mean equal to 0 and a standard deviation equal to 1. For the column Material, we used the Inverse Document Frequency (IDF) technique, which converts a categorical column into a numerical column of positive values, based on the frequency of each level of the attribute [12]. The resulting preprocessed data resulted in a reduction of records from 14,614 to 14,101 production orders.

Modeling For the modeling phase, we selected four open-source AutoML tools detailed in Section 3.2: AutoGluon, H2O AutoML, rminer, and TPOT. The validation method we used was based on a recent AutoML benchmark [5], adopting a 10-fold external cross-validation. Additionally, for each AutoML training, we used an internal 5-fold cross-validation to select the best algorithm and hyperparameters, a task that is performed automatically by these tools.

The test set predictions for each of the external 10 folds were evaluated using the Mean Absolute Error (MAE) ($\in [0.0,\infty[$, where the lower the value, the better are the predictions). In addition, we computed the Normalized MAE (NMAE), which is equal to the MAE divided by the range of the target column (in percentage). For the internal validation (performed by the AutoML tools), we used the MAE. To keep the comparison fair between the AutoML tools, we defined a maximum training of one hour when available (only unavailable on rminer).

Evaluation To evaluate the predictive results we computed the average metrics of the external 10 folds. We also added a confidence interval based on the t-distribution with 95% confidence, to ensure the statistical significance of the experiments. Table 3 shows the average results for the first CRISP-DM iteration. For each tool, the table shows the algorithm that was most often the leader during training across the 10 folds (**Best Algorithm**). The results also show the average test scores of the external 10 folds and confidence intervals (**MAE** and **NMAE**), with the best scores in bold.

Table 3. Average results obtained on the first CRISP-DM iteration on test data (best values in **bold**).

AutoML Tool	Best Algorithm	MAE	NMAE
AutoGluon	Ensemble	4.00±0.15	3.46±0.26
H2O AutoML	Stacked Ensemble	**3.70±0.10**	**3.21±0.24**
rminer	Stacked Ensemble	3.98±0.12	3.45±0.27
TPOT	Gradient Boosting	3.77±0.12	3.26±0.25

The results show that H2O AutoML obtained the best results with an average error of 3.70 days, corresponding to an average NMAE of 3.21%. Nevertheless, the four tools obtained close results (average MAE ranging from 3.70 to 4.00 and average NMAE ranging from 3.21% to 3.46%). After this iteration, the company provided an additional dataset related to the individual operations of a production order, which were addressed in the second CRISP-DM iteration.

4.2 Second CRISP-DM Iteration

Business Understanding During a new phase of business understanding, the company provided new insights about the manufacturing process. Each production order is composed of a series of operations (e.g., cut, paint). Each one of these operations is carried out in a specific work center and might involve different types of raw materials. Also, every production operation is associated with a production order ID and has a defined start and end timestamp. In this CRISP-DM iteration, we attempt to improve the previous iteration results by predicting the time needed to carry out production operations (instead of production orders). By predicting this attribute, we can estimate the total time needed to conclude the associated production order.

Data Understanding For this phase, a new data source was analyzed (Operations), related to production operations associated with a specific production order. The dataset has 40,610 records and 8 attributes. Each production order has between one and nine operations. The most common number of operations is three, while the average value is 2.83. Table 4 summarizes the attributes of the operations dataset.

Table 4. Description of the Production Operations dataset attributes.

Attribute	Description	Type	Levels	Example
Order_ID	ID of the production order	Categorical	14438	1000234997
Operation_ID	ID of the operation type (e.g., cut)	Categorical	36	25
Work_Center	Work center that performed the operation	Categorical	21	1LE02
Operation_quantity	Quantity associated to the operation	Numerical	-	44587
Start_date_op	Starting date of operation	Date	385	16/09/2020
End_date_op	Ending date of operation	Date	390	04/11/2020
Start_time_op	Starting time of operation	Time	29362	10:54:27
End_time_op	Ending time of operation	Time	29369	14:26:02

Data Preparation The creation of the target attribute followed an approach similar to the first iteration. We created a target column that specifies the number of days that an operation takes to finish. In this case, since the operations dataset has start and end times, we computed the difference between the starting timestamp and the ending timestamp. The result is that the values of target variable were converted to decimals (e.g., 1.25 days). For the data preprocessing we also used similar transformations. First, we deleted rows with missing values. Then, we used the same threshold to remove outliers from the target, so that operations that took more than 150 days were removed from the dataset (only 12 records were discarded).

Given that several production operations are associated with the same production order, we did not need to remove duplicate Order IDs. However, we removed operations associated with Order IDs that were not on the Orders dataset (used on the first CRISP-DM iteration). Finally, we standardized the Operation_quantity column and applied IDF to the columns Order_ID, Operation_ID, and Work_Center. The Data Preparation phase reduced the number of records from 40,610 to 39,225.

Modeling The modeling phase was similar to the first iteration, detailed in Section 4.1. It used the same four AutoML tools (AutoGluon, H2O AutoML, rminer, and TPOT), the same validation (10-fold external cross-validation and 5-fold internal cross-validation), as well as the same predictive metrics (MAE and NMAE).

Evaluation Given that in this iteration we predict the operation time instead of the total production time, we aggregate the Operations dataset after the predictions. We group the dataset by Order_ID and sum the time for all individual operations, resulting in a predicted production time. Then, we compare the obtained production time with the real values from the Orders dataset. Thus, instead of measuring the MAE and NMAE for the individual operations, we aggregate the results to maintain a meaningful comparison between CRISP-DM iterations. Also, the main problem identified by the company is related to the total production time instead of operations time. Table 5 shows the average results for the second CRISP-DM iteration.

Table 5. Average results obtained on the second CRISP-DM iteration on test data (best values in **bold**).

AutoML Tool	Best Algorithm	MAE	NMAE
AutoGluon	Ensemble	6.00±0.10	4.46±0.35
H2O AutoML	Deep Learning	**5.83±0.68**	**4.35±0.76**
rminer	Suppor Vector Machine (SVM)	6.51±0.11	4.82±0.38
TPOT	Gradient Boosting	6.64±0.11	4.92±0.40

The second iteration results show that, in general, this approach obtained worse results, when compared to the first iteration. Even though the best AutoML tool for this iteration (H2O AutoML) obtained an average error of 5.83 days, this is still worse than all the results from the first iteration (which had an average MAE ranging from 3.70 to 4.00 days).

These results suggest that, in this case, predicting production time by predicting each operation time is less effective than using the ML models to directly predict the total production time. After analyzing the results of this iteration with the company experts, we concluded that the degradation of the results was caused by two factors: first some operations may be executed simultaneously in two or more production lines, and the second was that different operations have different machine setup times, which the company does not store. Instead, it uses the average time for all operations. To check if we could improve the results, we decided to perform a third CRISP-DM iteration, using both the Orders and Operations dataset together.

4.3 Third CRISP-DM Iteration

On this iteration, we used the two datasets from the previous iterations (Orders and Operations). We merged the two datasets so that we remain with one record for each Order ID, but with additional columns regarding which operations are performed for each order. Since we used the same datasets, the business understanding and data understanding phases were skipped for this iteration.

Data Preparation In order to merge the two datasets (Orders and Operations), we first removed the records that had missing values from each dataset. Then, we pivoted the Operations dataset, such that each different operation became a new column. Since this dataset has 36 different operations (as shown on Table 4), this created 36 new columns for the Operations dataset. After the pivot, we changed the values with 1 and 0 (1, if the order includes the operation; 0, otherwise). Next, we performed a left join between the Orders and Operations. This join added 36 new columns to the Orders dataset (those created by the Operations pivot). Then, we created the target column by subtracting the end date from the initial date. Fig. 1 exemplifies the merging process.

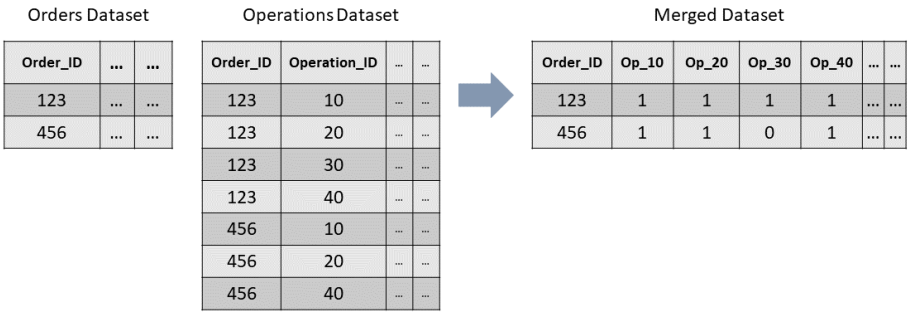

Fig. 1. Representation of the merging process.

Finally, similarly to the first iteration, we standardized the Order_quantity column and applied IDF to the Material column. This resulted in a dataset with 14,106 rows and 39 features (excluding the target column).

Modeling The modeling phase was similar to the previous two CRISP-DM iterations, using AutoML (Section 4.1).

Evaluation Table 6 shows the average results for the third CRISP-DM iteration, including, for each AutoML tool, the algorithm that was most often the leader, the average predictive scores (MAE and NMAE), and confidence intervals (t-distribution with 95% confidence).

Table 6. Average results obtained on the third CRISP-DM iteration on test data (best values in **bold**).

AutoML Tool	Best Algorithm	MAE	NMAE
AutoGluon	Ensemble	3.38±0.07	2.81±0.31
H2O AutoML	Stacked Ensemble	**3.03±0.04**	**2.53±0.29**
rminer	Stacked Ensemble	3.42±0.08	2.85±0.31
TPOT	Gradient Boosting	3.17±0.08	2.64±0.29

The results from the third iteration show that for the third time H2O AutoML obtained the best results, with an average error of 3.03 days, corresponding to 2.53% of the target range. Similar to the first iteration, all tools obtained close results, with a maximum difference of 0.39 for MAE and 0.32 percentage points for NMAE. Overall, this was the iteration that achieved the best predictions.

5 Discussion

Table 7 presents a summary of the predictive results for the three CRISP-DM iterations. For each **Iteration**, we detail the used datasets (**Datasets**). For each **AutoML Tool**, we also include the algorithm that was most often the leader during training across the 10 folds (**Best Algorithm**). The results show the average test scores of the external 10 folds (**MAE** and **NMAE**), with the best values for each iteration in bold.

Table 7. Average results obtained on each CRISP-DM iteration on test data (best values in **bold**).

Iteration	Datasets	AutoML Tool	Best Algorithm	MAE	NMAE
1st	Orders	AutoGluon	Ensemble	4.00±0.15	3.46±0.26
		H2O AutoML	Stacked Ensemble	**3.70±0.10**	**3.21±0.24**
		rminer	Stacked Ensemble	3.98±0.12	3.45±0.27
		TPOT	Gradient Boosting	3.77±0.12	3.26±0.25
2nd	Operations	AutoGluon	Ensemble	6.00±0.10	4.46±0.35
		H2O AutoML	Deep Learning	**5.83±0.68**	**4.35±0.76**
		rminer	SVM	6.51±0.11	4.82±0.38
		TPOT	Gradient Boosting	6.64±0.11	4.92±0.40
3rd	Orders and Operations	AutoGluon	Ensemble	3.38±0.07	2.81±0.31
		H2O AutoML	Stacked Ensemble	**3.03±0.04**	**2.53±0.29**
		rminer	Stacked Ensemble	3.42±0.08	2.85±0.31
		TPOT	Gradient Boosting	3.17±0.08	2.64±0.29

The results became worse between the first and second iterations, when we tried to predict the production time using only data from operations. While in the first iteration the best AutoML tool (H2O AutoML) obtained an average MAE of 3.70 days, the best results in the second iteration were almost doubled (MAE of 5.83 days, also by H2O AutoML). As mentioned in Section 4.2, the degradation of results obtained during the second iteration can be related to the absence of information about setup time and parallel operations.

On the third iteration, when we merged the Orders and Operation datasets, the predictive results presented an improvement from both the first and second iterations. In effect, the third iteration obtained the best predictive results, with the best MAE of 3.03 days and NMAE of 2.53%, by H2O AutoML. Nevertheless, on the third iteration, all tools achieved average results better than all the other results from the first and second iterations. This suggests that using data about individual operations might be useful for improving production time for the whole order.

It is worth mentioning that, on all three CRISP-DM iterations, the AutoML results were close: maximum MAE difference of 0.30 for 1st iteration; 0.81 - 2nd iteration; 0.39 - 3rd iteration. For the NMAE, the maximum predictive differences

were: 0.25 *percentage points (pp)* for the 1st iteration; 0.57 *pp* - 2nd iteration; 0.32 *pp* - 3rd iteration. Regarding the best performing AutoML tools, H2O AutoML achieved the best results on all three iterations. TPOT was the second-best tool on iterations 1 and 3. AutoGluon was the second-best tool on the 2nd iteration.

6 Conclusion

In this paper, we applied the CRISP-DM methodology to predict production time for the lithography process of a Portuguese manufacturing company. This company produces metal cans for several sectors (e.g., food, cosmetics, pharmaceutical) and each type of can might involve different types of operations (e.g., cutting, applying varnish, printing). The prediction of the production time is essential for the company, in order to further optimize the production plan.

We executed a total of three CRISP-DM executions, using data provided by the company, between December 2018 to December 2020. The company provided two datasets: the first one (Orders) included data about production orders; the second one (Operations) provided information about individual operations associated with production orders. For the modeling phase of CRISP-DM, we used AutoML to automatically select the best ML algorithms and hyperparameters. We used four recent open-source AutoML tools: AutoGluon, H2O AutoML, rminer, and TPOT. During the first CRISP-DM iteration, we only used data from the Orders dataset. For the second iteration, we predicted production time using only the Operations dataset. In the third and final iteration, we merged the Orders and Operations datasets. The best predictive results were achieved on the third CRISP-DM iteration (average MAE of 3.03 days, by H2O AutoML).

In future work, we intend to explore more AutoML technologies to automate the Modeling phase of CRISP-DM. Moreover, we intend to address issues related to the second iteration, namely the setup time and usage of parallel operations. Lastly, we wish to use more datasets from the analyzed manufacturing domain to predict production time, aiming to check if there is further consistency in the results obtained in this paper.

Acknowledgements

This work has been supported by the European Regional Development Fund (FEDER) through a grant of the Operational Programme for Competitivity and Internationalization of Portugal 2020 Partnership Agreement (POCI-01-0247-FEDER-046102, PRODUTECH4S&C).

References

1. Asadzadeh, S., Azadeh, A., Ziaeifar, A.: A neuro-fuzzy-regression algorithm for improved prediction of manufacturing lead time with machine breakdowns. Concurrent Engineering **19**(4), 269–281 (2011)

2. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al.: Crisp-dm 1.0: Step-by-step data mining guide. SPSS inc **9**, 13 (2000)
3. Cortez, P.: Data mining with neural networks and support vector machines using the r/rminer tool. In: Advances in Data Mining. vol. 6171, pp. 572–583. Springer (2010). https://doi.org/10.1007/978-3-642-14400-4_44
4. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A.: Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505 (2020)
5. Ferreira, L., Pilastri, A.L., Martins, C.M., Pires, P.M., Cortez, P.: A comparison of automl tools for machine learning, deep learning and xgboost. In: International Joint Conference on Neural Networks, IJCNN 2021. pp. 1–8. IEEE (2021). https://doi.org/10.1109/IJCNN52387.2021.9534091
6. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Advances in Neural Information Processing Systems. vol. 28 (2015)
7. Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S.H., Rodríguez-Roda, I., Holmes, G.: Which method to use? an assessment of data mining methods in environmental data science. Environ. Model. Softw. **110**, 3–27 (2018). https://doi.org/10.1016/j.envsoft.2018.09.021
8. Gyulai, D., Pfeiffer, A., Nick, G., Gallina, V., Sihn, W., Monostori, L.: Lead time prediction in a flow-shop environment with analytical and machine learning approaches. IFAC-PapersOnLine **51**(11), 1029–1034 (2018). https://doi.org/10.1016/j.ifacol.2018.08.472
9. Karaoglan, A.D., Karademir, O.: Flow time and product cost estimation by using an artificial neural network (ann): A case study for transformer orders. The Engineering Economist **62**(3), 272–292 (2017)
10. LeDell, E., Poirier, S.: H2O AutoML: Scalable automatic machine learning. 7th ICML Workshop on Automated Machine Learning (AutoML) (2020)
11. Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., Sihn, W., Monostori, L.: Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. Procedia CIRP **72**, 1051–1056 (2018). https://doi.org/10.1016/j.procir.2018.03.148
12. Matos, L.M., Cortez, P., Mendes, R., Moreau, A.: A comparison of data-driven approaches for mobile marketing user conversion prediction. In: 9th IEEE International Conference on Intelligent Systems, IS 2018. pp. 140–146. IEEE (2018). https://doi.org/10.1109/IS.2018.8710472
13. Okoshi, C.Y., Pinheiro de Lima, E., Gouvea Da Costa, S.E.: Performance cause and effect studies: Analyzing high performance manufacturing companies. International Journal of Production Economics **210**, 27–41 (2019). https://doi.org/10.1016/j.ijpe.2019.01.003
14. Olson, R.S., Moore, J.H.: TPOT: A tree-based pipeline optimization tool for automating machine learning. In: Proceedings of the 2016 Workshop on Automatic Machine Learning. vol. 64, pp. 66–74. JMLR.org (2016)
15. Purbasari, A., Rinawan, F.R., Zulianto, A., Susanti, A.I., Komara, H.: Crisp-dm for data quality improvement to support machine learning of stunting prediction in infants and toddlers. In: ICAICTA 2021. pp. 1–6 (2021)
16. Reuter, C., Brambring, F.: Improving data consistency in production control. Procedia CIRP **41**, 51–56 (2016)
17. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. vol. 1, pp. 29–40. Manchester (2000)