Applied Computer Science, vol. 19, no. 3, pp. 83–95
doi: 10.35784/acs-2023-26

Submitted: 2023-09-10 | Revised: 2023-09-14 | Accepted: 2023-09-20

Keywords: data engineering, data mining, CRISP-DM, assembly, process planning

Jolanta BRZOZOWSKA [0000-0002-4355-2847]*, *Jakub PIZOŃ* [0000-0002-0806-6771]**
Gulzhan BAYTIKENOVA[***], *Arkadiusz GOLA* [0000-0002-2935-5003]****,
Alfiya ZAKIMOVA[***], *Katarzyna PIOTROWSKA* [0000-0003-0899-7610]****

DATA ENGINEERING IN CRISP-DM PROCESS PRODUCTION DATA – CASE STUDY

Abstract

The paper describes one of the methods of data acquisition in data mining models used to support decision-making. The study presents the possibilities of data collection using the phases of the CRISP-DM model for an organization and presents the possibility of adapting the model for analysis and management in the decision-making process. The first three phases of implementing the CRISP-DM model are described using data from an enterprise with small batch production as an example. The paper presents the CRISP-DM based model for data mining in the process of predicting assembly cycle time. The developed solution has been evaluated using real industrial data and will be a part of methodology that allows to estimate the assembly time of a finished product at the quotation stage, i.e., without the detailed technology of the product being known.

1. INTRODUCTION

Processing massive amounts of data to support decision-making processes is becoming increasingly important in corporate strategies (Krcmar & Helmut, 2015; Laudon et al., 2010). Recent large database mining projects are by far dominated by the CRISP-DM methodology (Shearer, 2000) developed by MIT (42% of applications). In the second place there are proprietary methodologies (19%) and in third place the methodology SEMMA proposed by SAS (13%) (Rohanizadeh & Moghadam, 2009; Moutinho & Huarng, 2015). The CRISP-DM model describes in detail how to collect and analyze data that can be

[*] Lublin University of Technology, Faculty of Mechanical Engineering, Department of Production Computerisation and Robotisation, d562@pollub.edu.pl
[**] Lublin University of Technology, Faculty of Management, Department of Enterprise Organization, j.pizon@pollub.pl
[***] School of Business and Entrepreneurship, D. Serikbayev East Kazakhstan Technical University, Kazakhstan, gbaytikenova@mail.ru, azakimova@edu.ektu.kz
[****] Lublin University of Technology, Faculty of Mechanical Engineering, Department of Production Computerisation and Robotisation, a.gola@pollub.pl, k.piotrowska@pollub.pl

further processed and build models from it for corporate decision-making. CRISP-DM allows the creation of a data mining model tailored to specific needs.

This paper is developed using the CRISP-DM (Cross Industry Process Model for Data Mining) methodology that follows a goal-oriented approach. It is a mature approach that continues to be widely accepted in data mining projects through data machine learning mining algorithms (Ayele, 2020). This methodology provides a life cycle approach in methodology for the knowledge discovery process in databases (Martinez-Plumed, 2019).

First, information on business objectives was collected, which is used in the data mining process. The CRISP-DM methodology allows this task to be carried out in a structured and transparent way. The next very important step was to understand the data and select qualitative data. This step is crucial - it helps to prevent unexpected problems in the next phase. The data preparation phase was the longest and most labor-intensive phase of the CRISP-DM process so far. Subsequent phases will systematically follow the phases described in the CRISP-DM model (Cheng, 2023).

Data for the model was obtained from a small-batch production company that assembles machines. The company's production process is carried out in production cells. The assembly of the machine begins with the frame, on which, using an overhead crane and a forklift, the mechanical elements, pneumatics, electrical board, and cabinet are successively mounted. In the final stage, covers are installed.

A research problem that has arisen is the lack of ability to predict machine assembly times at the quotation stage, a stage at which detailed assembly technology has not yet been developed. The aim of this paper is to present the concept of an intelligent system for the prediction of assembly times of complex products using artificial neural networks. In particular, the main input and output factors for predicting the duration of the assembly process are identified and analysed.

2. BUILDING OF DATA MINING MODELS

Data mining provides direct benefits in business decisions (Smyth, Hand & Mannila, 2001; Nisbet, Elder & Miner, 2009). The process of building a decision model in data mining consists of four major steps (Hastie, Tibshirani & Friedman, 2001):

- initial exploration,
- model building with pattern identification,
- model evaluation and verification,
- implementation and application of the model to new data to obtain predicted values or classifications.

Exploration is the initial stage of model building, which begins with data preparation (Surma, 2009). It mainly includes cleaning and transformation, separation of subsets of records and selection of data attributes, the purpose of which is to reduce the number of analyzed variables to a level that allows us to effectively perform analysis (this level depends on the data mining methods used). Once the data is prepared, the further course of exploration depends on the specific problem we want to solve. Exploration can involve a wide variety of methods, from simple selection of predictors using linear regression to sophisticated examination of the data using various graphical and statistical methods, the purpose of which is to select the most important features and determine the overall nature

and complexity of the model for the second stage of data mining. To conduct successful data mining projects, in many cases, interdisciplinary collaboration between different domain experts is required (Choudhary, Harding & Popplewell, 2006).

In the model building and evaluation stage, various models are considered, after which the best one is selected. The evaluation criterion is the quality of prediction, that is, the correctness of determining the value of the modeled variable and the stability of the results for different samples. At first glance, the selection of the best model may seem a fairly simple task, but in practice, it is sometimes a complicated process. There are many different methods for evaluating models and selecting the best one. Often techniques based on comparative evaluation of models (competitive evaluation of models) are used, which involves applying individual methods to the same data sets and then selecting the best one or building a composite model (Zaskórski & Pałka, 2012).

The techniques for evaluating and combining models, which are a key part in data mining, boil down to the following:

- aggregation of models (voting and averaging - bagging),
- model amplification (adaptive sampling),
- combining models (boosting),
- generalizing models (stacking, stacked generalizations),
- learning models (meta-learning).

The final stage of building a data mining model is implementation and application, in which the model produced and found to be the most suitable model is applied to the new data. The finished model is applied to obtain predicted values or classifications (Sturm, 2000).

Quite an important concept used when building a data mining model is model aggregation (bagging), as well as voting (voting) and veraging (veraging). Model aggregation involves a method of predicting multiple models of the same type obtained for different learning sets or multiple models of different types obtained for the same data set. Modeling in this case of a continuous variable creates a procedure called averaging, and in the case of qualitative variables on classification undertakings, constitutes voting. The use of model aggregation enables more accurate and reliable results for complex relationships. It is also used to solve the problem of instability and small discrepancies in the results obtained when using a complex method for a small data set. In case of highly divergent results, the model amplification method can be used.

Model amplification (boosting) is used to build successive models for the data and determine the weights for the main model. The first model is built with the same weights for all cases, and in subsequent stages the case weights are modified to obtain more accurate predictions for those cases for which earlier models gave erroneous predictions. The amplification also makes it possible to create a sequence of models, each of which is patterned in making predictions for cases that the preceding models failed to handle.

Data preparation and verification is extremely important in the data mining process. Analyzing excessive and redundant data without solving the above problems results in misleading results, especially during the creation of a data mining model used to predict various activities and events, or the future. Data reduction in data mining refers to activities aimed at aggregating data into a form that is easier to perceive and process (Han, Kamber & Pei, 2011). Simple tabular techniques, descriptive statistics and more

sophisticated techniques such as cluster and principal component analysis are used to reduce data.

Implementation and application in data mining means applying the results of the analysis to new data. It is used in so-called predictive data mining and in model-free classification. Once a satisfactory model or segmentation is obtained, these results should be applied so that predicted values or segment memberships can be quickly obtained. A very popular method of model building is the drill-down analysis technique. In data mining, it involves an interactive examination of data, most often for large databases. The process of data drill-down begins by performing simple cross sections against several variables (time, distance, region). A variety of statistics, summaries and summaries are determined for each group. At the lowest level, referred to as the bottom, we have access to elementary data, which is the primary source of power for the model.

In data mining, the term "machine learning" is often used, understood as a general term for model fitting algorithms. Unlike traditional statistical data analysis, in which we estimate population parameters using statistical methods, in machine learning data mining the emphasis is on the accuracy and utility of the predictions, or on the accurate description of the resulting data. Meta-learning is also used in predictive data mining models to combine the results of multiple models into a single generalized model. This technique is particularly useful when the models are of different types (Frawley, Piatetsky-Shapiro & Matheus, 1998). This method is often applied to the results of a pooled model, obtained by meta-study, which can be applied repeatedly. However, in practice, it increases the amount of computation, and the obtained improvement of models is less and less significant. The implementation of complex data mining (data mining) projects in business organizations requires the coordinated efforts of experts, specialists and analysts from different departments of the organization.

In order to achieve the expected result, a specific methodology is required, which can serve as a scenario for how the process of collecting and analyzing data, disseminating results and checking the benefits of implementing the model project should be organized. One of the main methods for building a data mining model is the CRISP (Cross-Industry Standard Process for Data Mining) method. This method has become a widely available standard for the data mining process (Gröger, Niedermann & Mitschang, 2012).

3. THE CHARACTERISTICS OF CRISP-DM

CRISP-DM is an industry-independent data mining process model. The CRISP-DM approach offers many advantages in terms of data preparation, modeling and evaluation steps (Weller, Roesmann, Eggert, von Enzberg, Gräßler, Dumitrescu, 2023). It consists of six iterative stages, from business understanding to implementation. Figure 1 shows the phases of the data mining process.

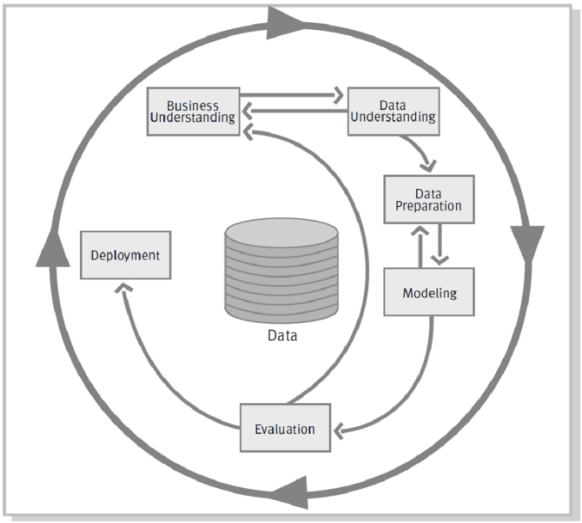

Fig. 1. Phases of the CRISP-DM reference model (Hubera et al., 2018)

The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase or which specific phase task should be performed next. The arrows indicate the most important and frequent dependencies between phases. The outer circle in Figure 1 symbolizes the cyclical nature of data mining itself. Data mining is not over once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones (Chapman et al., 2000; Hubera et al., 2018; Santos & Azevedo, 2005).

In the following, we outline each phase briefly (Chapman et. Al. 2000; Schröerab, Kruseb & Gómezb, 2020):

1) **Business understanding –** This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.
2) **Data understanding –** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3) **Data preparation –** The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as

transformation and cleaning of data for modeling tools. Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent.

4) **Modeling –** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones.
5) **Evaluation –** At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
6) **Deployment –** Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision-making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carriedout in order to actually make use of the created models.

Figure 2 summarizes the main inputs to the deliverables. This does not mean that only the inputs listed should be considered – for example, the business objectives should be pervasive to all deliverables. However, the deliverables should address specific issues raised by their inputs.

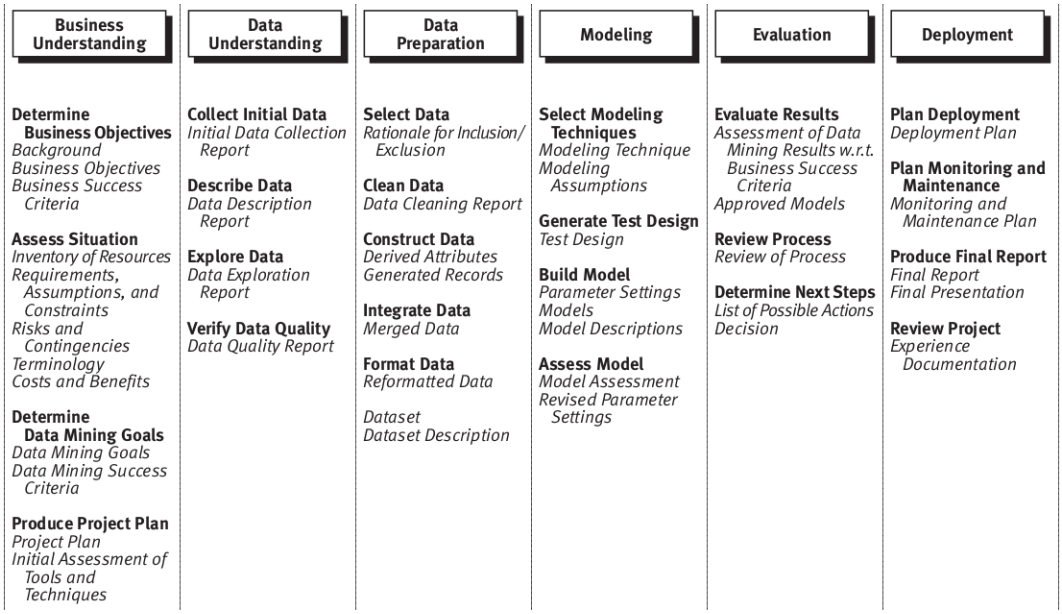

Fig. 2. CRISP-DM reference model [Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, Wirth 2000]

4. BUSINESS UNDERSTANTING / DATA UNDERSTANTING

The purpose of data mining is to seek an answer to the question of how to estimate the assembly time of a finished product at the quotation stage, i.e., without the detailed technology of the product being known. In the case of a large amount of data, finding the optimal solution to the problem requires analyzing all the data related to the assembly of the machine and finding relationships between related data. An artificial neural network will be helpful here, which can be used to determine and estimate time in the assembly process planning (Brzozowska & Gola, 2021). The development of the model with the use of ANN can be based on the following steps: development of training and test sets and finding the best SANN structure.

Preliminary results conducted on simulation data helped to determine what input factors should be considered for the model. The assembly process was analyzed and the features (attributes) that affect the time norm were selected (Fig. 3). By inputting into the model the inputs of part availability, resource availability and novelty factor after running the model, a predicted/estimated machine assembly time will be generated at the output of the model.

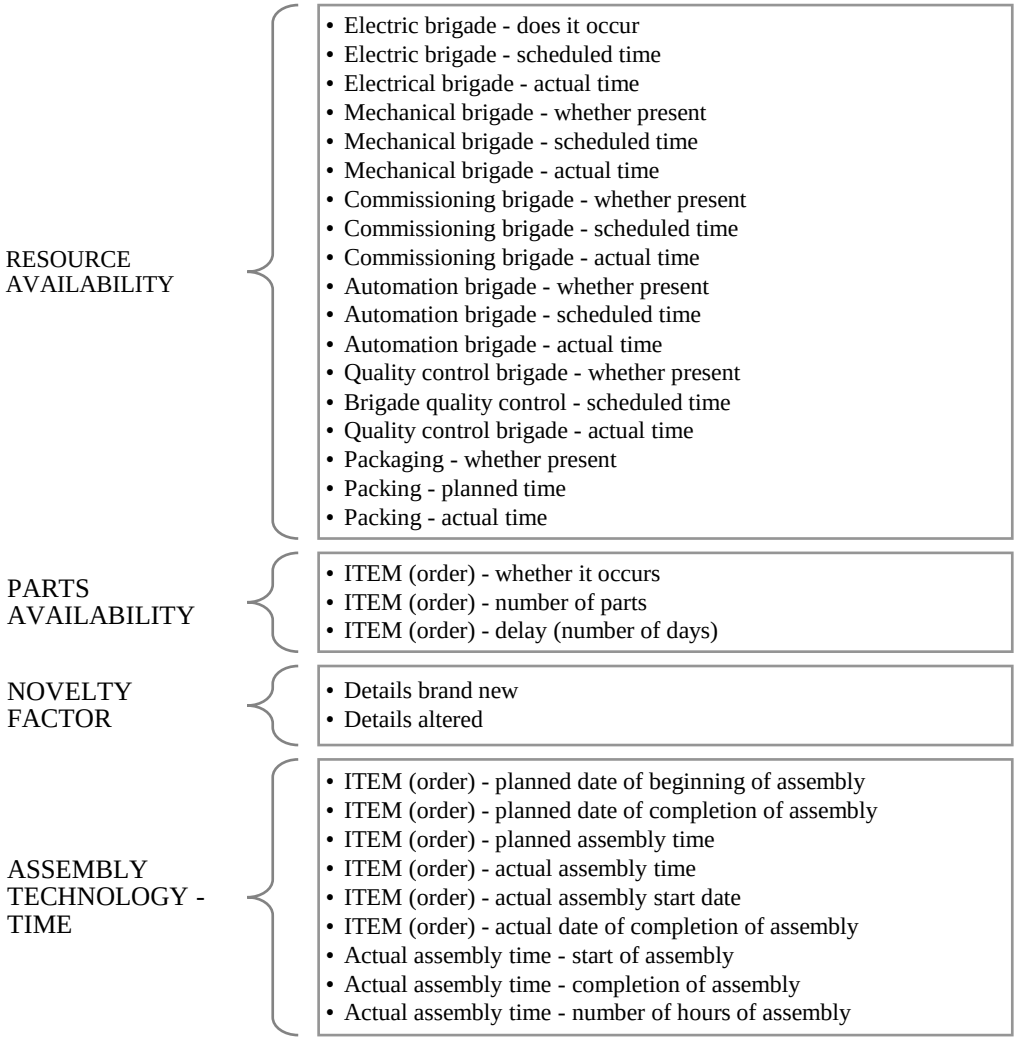

Fig. 3. Inputs to the model

5. DATA PREPARATION

In view of the previous steps of the CRISP-DM process, in which both the business case and the data representing that objective were identified, the next step is to perform data preparation.

The data for the survey process was extracted and prepared through an ETL (Extract Transfer Load) process. In this process, the first step was to export data from the level of the company's domain systems. As part of this step, the native functionalities of the information systems solutions were used. Data was extracted from both ERP (Enterprise Business Planning) and BI (Business Intelligence) reporting systems. In addition, flat files were also exported as a direct source of reporting at the operational level.

In order to unify the sources, tabular files in xlsx format were chosen as the output format. In view of this, multiple sources of different data were obtained. In the next step, an analysis of the meaning of the designations of successive fields within the acquired data tables was carried out. The purpose of the analysis was to interpret and label the columns adequately for modeling purposes. At the same time, the usefulness of individual indicators was assessed. As a result, field designations were supplemented as well as redundant data or redundant ones were removed.

The next step was to select files so that they express the planned data model. The data model directly responds to the needs described in business terms. The data were selected in detail and described. In view of this, on the basis of the criterion of data belonging to the model, those files and those fields were selected that will be used to draw up the target research model.

Due to the size of the source files as well as the number of instances of cases represented, it was decided that the data model preparation work would be performed using the pandas library implemented in Python. The analyses were carried out from the Anaconda environment.

In order to maintain consistency - according to the adopted model - the data was integrated at the level of the project number and adequately the production order number. Further adequately, inversely to the normalization process, further ranges of data were integrated, expanding the model with the necessary dimensions. So that, as a result, the collected data represented the desired range, i.e.: resource availability, part availability, novelty factor and assembly technology - time (fig. 4).

```
In [1]: import pandas as pd
        df = pd.read_excel('1. Estimated vs Actual Hours Costs by Order,Work Center.xlsx', sheet_name = "data")

In [21]: df2 = pd.read_excel('5. 6. Hours , Hours Project.xlsx', sheet_name = "data")

In [ ]: df3 = pd.read_excel('8.9. Production Planning by Order,Task, PO Intended vs Verified Date', sheet_name = "data")
```

Fig. 4. Data import into the environment

An important step in implementing .xlsx type files into the pandas library is to prepare them. All columns should be described as clearly as possible. Missing data in the cells are marked as NaN.

Figure 5 shows a fragment of an excel file and its implementation into the pandas library (fig. 6).

Import Status	Import Code	Import Message	Company	Production Order	Production Order Status	Project (PCS)	Item	Item Type Work	Item Type Work Desc	Actual Machine Costs (Aggregated)	Actual Overhead Costs (Aggregated)
			101	000111552	Closed		X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	0	0
			101	000111552	Closed		X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	0	0
			101	000111552	Closed		X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	0	0
			101	000111895	Closed		X_920024_INSPECTION	Job Shop	FINAL PRODUCT	0	2212,16
			101	000111895	Closed		X_920024_INSPECTION	Job Shop	FINAL PRODUCT	0	594,26
			101	000112077	Closed		X_304181_0800-04000	Job Shop	packing	0	0
			101	000112079	Closed		X_304145_0800-04000	Job Shop	packing	0	0
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	684,25
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	1164,15
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	25346,65
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	0
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	3837,93
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	0
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	0
			101	000114089	Closed		X_920040_RODS PROCESSING	Job Shop	RODS PROCESSING UNITS	0	0

Fig. 5. An excerpt from the excel file implemented into the pandas library

```
In [ ]: df.select_dtypes

In [ ]: df.plot()

In [23]: df.head()
```

Out[23]:

	Import Status	Import Code	Import Message	Company	Production Order	Production Order Status	Project (PCS)	Item	Item Type Work	Item Type Work Desc	...	Actual Machine Costs (Aggregated)	Actual Overhead Costs (Aggregated)	A Subcontra C (Aggreg
0	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	0	0.00	
1	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	0	0.00	
2	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	0	0.00	
3	NaN	NaN	NaN	101	111895	Closed	NaN	X_920024_INSPECTION	Job Shop	FINAL PRODUCT	...	0	2212.16	

Fig. 6. Implemented excel file for pandas library

The entire ETL process was using the following python commands. At first, the flat file data was implemented as a DataFrame. In the next stopper from the merge command, a left join was performed to successively expand the scope of the data.

To be consistent - in accordance with the adopted model - the data was integrated at the level of the project number and, correspondingly, the production order number (fig. 7).

```
In [25]: production_data = pd.merge(df, df2, how = "left", on = "Production Order")

In [26]: production_data1 = pd.merge(production_data, df3, how = "left", on = "Production Order")
```

Fig. 7. Joining data on production order

The result was a file representing the desired data model, consisting of with 583129 rows and 81 columns (fig. 8). After cleaning, organizing and completion work, the data was exported to a csv format file. A fully described and integrated data model was obtained, which will be used for further analysis using artificial intelligence methods.

Out[26]:

	Import Status_x	Import Code_x	Import Message_x	Company_x	Production Order	Production Order Status	Project (PCS)	Item	Item Type Work	Item Type Work Desc	...	End Time	Unnamed: 29_y	Perc Con
0	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	NaN	NaN	
1	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	NaN	NaN	
2	NaN	NaN	NaN	101	111552	Closed	NaN	X_304145_26-061142	Job Shop	BUFFER FEEDER 800W	...	NaN	NaN	
3	NaN	NaN	NaN	101	111895	Closed	NaN	X_920024_INSPECTION	Job Shop	FINAL PRODUCT	...	Wednesday	826.0	
4	NaN	NaN	NaN	101	111895	Closed	NaN	X_920024_INSPECTION	Job Shop	FINAL PRODUCT	...	Wednesday	826.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
583124	NaN	NaN	NaN	101	420000282	Active	NaN	28-ITM-019873	Job Shop	TPSM 21993003 overhaul	...	Thursday	1630.0	
583125	NaN	NaN	NaN	101	420000282	Active	NaN	28-ITM-019873	Job Shop	TPSM 21993003 overhaul	...	Friday	1530.0	
583126	NaN	NaN	NaN	101	420000283	Planned	NaN	RM0940000-0036_EL_PARTS	Job Shop	SCTM - EL_PARTS	...	NaN	NaN	
583127	NaN	NaN	NaN	101	420000283	Planned	NaN	RM0940000-0036_EL_PARTS	Job Shop	SCTM - EL_PARTS	...	NaN	NaN	
583128	NaN	NaN	NaN	101	420000284	Planned	NaN	RM0940000-0036_EL_PLATE	Job Shop	SCTM - EL_PLATE	...	NaN	NaN	

583129 rows × 81 columns

Fig. 8. Output data model

6. SUMMARY AND CONCLUSIONS

Working according to the CRISP-DM model, we can achieve good quality model results. By going through the stages of business understanding, data understanding and data preparation, we were able to analyze the relevant data for building the model in the next steps. Such a basis will help in testing various artificial neural network models. The research carried out so far allows us to conclude that it is possible to develop a model using neural networks, which, after entering the input parameters, will generate information on the output about the estimated assembly cycle of the machine. The indicated model can be used to support analytical, planning and decision-making processes. Hence, identifying decision-making processes and improving the accuracy of decision-making affects the smooth operation of any organization by successively making appropriate changes to the model and observing their impact.

The development of a model using SSN can be based on the following steps: developing training sets and test sets and finding the best SSN structure. The developed method will be able to be used in enterprises as an intelligent system to support efficient and accurate estimation of the assembly time of machines not yet ordered. This will allow to increase the accuracy of enterprises' work, claims in meeting given production completion dates, and will increase competitiveness in the market. The system will be ready for implementation in production conditions for small batch and unit production.

Author contribution

Conceptualization, A.G.; methodology, J.B., J.P., A.G.; software, J.P.; validation, A.G. and J.P.; formal analysis A.G.; investigation, J.B., J.P., A.G.; resources, J.B..; data curation, J.B. and J.P.; writing—original draft preparation, J.B and J.P.; writing—review and editing, A.G. and K.P.; visualization, J.P.; supervision, A.G.; project administration, A.G.; funding acquisition, K.P.

Conflicts of Interest

The authors declare no conflicts of interests.

REFERENCES

Ayele, W.Y. (2020). Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Science and Applications*, *11*,(6), 20–32. https://doi.org/10.14569/IJACSA.2020.0110603

Brzozowska, J., Gola, A. (2021). Computer aided assembly planning using MS Excel software – a case study. *Applied Computer Science*, *17*(2), 70-89. https://doi.org/10.23743/acs-2021-14

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0. Step-by-step data mining guide. *SPSS*. https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf

Cheng, A. (2023), Evaluating Fintech insdustry's risks: A preliminary analysis based on CRISP-DM framework. *Finance Research Letters*, *55*(B), 103966. https://doi.org/10.1016/j.frl.2023.103966

Choudhary, A.K., Harding, J.A., Popplewell, K. (2006). Knowledge discovery for moderating collaborative projects. *4th IEEE International Conference on Industrial Informatics*, (pp. 519–524). IEEE. https://doi.org/10.1109/INDIN.2006.275610

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, *13*(2), 57. https://doi.org/10.1609/aimag.v13i3.1011

Gröger, C., Niedermann, F., & Mitschang B. (2012). Data mining-driven manufacturing process optimization. *World congress on engineering*, 14461305.

Han J., Kamber M., Pei J. (2011). Data Mining. Concepts and techniques, third edition, *The Morgan Kaufmann Series in Data Management Systems*, San Francisco, CA. https://doi.org/10.1016/C2009-0-61819-5

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning: Data mining, inference, and prediction, Second Edition, *Springer*, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Huber, S., Wiemer, H., Schneider, D., Ihlenfeldt, S. (2018). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM Model. *Procedia CIRP*, *79*, 403-408, https://doi.org/10.1016/j.procir.2019.02.106

Krcmar, H. (2015). Informationsmanagement. *Springer*, Berlin-Heidelberg. https://doi.org/10.1007/978-3-662-45863-1

Laudon, K.C., Laudon J.P., & Schoder D. (2010). Wirtschaftsinformatik: Eine Einführung. *Pearson*, München, Deutschland.

Martinez-Plumed F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., Flach, P. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories, *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048-3061. . https://doi.org/10.1109/TKDE.2019.2962680

Moutinho L., Huarng K.-H. (2015). Quantitative modelling in marketing and management, *World Scientific Publishing*, Singapore.

Nisbet, R., Elder, J., Miner G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier. https://doi.org/10.1016/B978-0-12-374765-5.X0001-0

Rohanizadeh, S.S., Moghadam, M.B. (2009). A Proposed Data Mining Methodology and its Application to Industrial Procedures, *Journal of Industrial Engineering*, 37-50.

Santos, M., Azevedo, C. (2005). Data Mining – Descoberta de Conhecimento em Bases de Dados. FCA Publisher, https://hdl.handle.net/1822/19136Schröer, C., Kruse, F., Gómez, J. C. M. (2021). A Systematic Literature Review of Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Warehousing*, *5*(4), 13-22.

Smyth, P., Hand, D., & Mannila, H. (2001). *Principles of Data Mining*, The MIT Press, 026208290x.

Sturm, J. (2000). *Hurtownie danych. SQL Server 7.0*, Przewodnik techniczny. APN PROMISE.

Surma, J. (2009). *Business Intelligence*. Systemy wspomagania decyzji biznesowych. *PWN*, Warsaw.

Weller, J., Roesmann, D., Eggert, S., Von Enzberg, S., Gräßler, I. &, Dumitrescu, R. (2023). Identification and prediction of standard times in machining for precision steel tubes through the usage of data analytics. *Procedia CIRP*, *119*, 514-520. https://doi.org/10.1016/j.procir.2023.01.011

Zaskórski, P., & Pałka, D. (2012). *Data mining in decision-making processes*. Warsaw School of Information Technology. Scientific Journals. 143-161.