IJIO
(*International Journal of Industrial Optimization*)
ISSN 2723-3022 (online) | 2714-6006 (printed)
Vol 6, No. 1, 87-102
https://doi.org/10.12928/ijio.v6i1.11357

Optimization of throughput rate prediction in animal feed industry using crisp-dm and operational research approaches

Riri Indah Lestari *, Luciana Andrawina, Ilma Mufidah

Industrial Engineering Department, Telkom University, Indonesia
* Corresponding Author: ririlestari@student.telkomuniversity.ac.id

ARTICLE INFO

ABSTRACT

Article history
Received: August 13, 2024
Revised: September 10, 2024
Accepted: October 23, 2024

Keywords
Throughput rates;
CRISP-DM;
Data mining;
Machine learning;
Operation Research.

The competitive animal feed industry requires efficient production planning to meet market demand, maximize resource use, and sustain profitability. Various raw materials, tools, and techniques are utilized to create animal feed, which results in various variants that might influence throughput rates and thereby alter the accuracy of yield projections. Data mining is applied to train and validate different algorithms to ascertain the most effective model for predicting throughput rates through machine learning. This study uses CRISP-DM to construct an enhanced predictive model for production throughput rate. Due to the model's improved prediction accuracy, scheduling and operational decision-making will be more efficient and cost-effective. The CRISP-DM framework is used to examine historical production data and forecast production levels. Advanced machine learning techniques train and evaluate the model to make accurate predictions that can be mathematically simulated using possible constraints. The findings show that throughput rate predictions are effectively generated by the predictive model that was created using data mining processes. Metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess the model and identify the optimal model after attempting using different predictive machine learning techniques. With the linear regression algorithm and MAE values of 5,186, MSE of 1,585, and RMSE of 5,970.32, the best prediction model test results have been determined. An optimal scheduling simulation is conducted from the selected model, with the constraint of the customer's delivery requirements and the time capacity, specifically a maximum throughput rate prediction of 23.78 tons/hour. However, this study reveals how the data mining process is applied to the decision-making process with the use of operation research support so that the optimal production rate prediction is 22 tons/hour.

1. Introduction

The subject of the study is an animal feed plant that produce four different kinds of products with different levels of complexity. That four types of feed products are differentiated based on their physical form—pellet, crumble, fine, and mash—all of which are produced using serial batch processing system. There are numerous strategies that can be implemented to address the challenges in the feed industry, such as the pursuit of alternative raw materials, developing innovative feed formulas, enhancement of production efficiency, and competition in the field of technology *(Article of Strategies for Facing Feed Industry Challenges, 2022).* Instead of processing one product continuously, batch production systems process several goods in different batch sizes, that requires the initial configuration of new equipment in between batches [1]. Currently, the batch serial production plan contributes to bottlenecks due to planned product changeovers, thus affects the achievement of throughput rates [2][3]. In a serial production system, where materials flow sequentially from one production station to the next and the set of production stations is organized in a sequential production process, varying processing times or planned or unplanned interruption events can affect production throughput [4]. Hiller et.al [5] developed a systematic approach to classify prediction models using machine learning (ML) in the production scheduling structure to predict throughput time for on-time customer delivery. While in this study, production throughput will be predicted using machine learning and data mining techniques based on a dynamic production plan. This is based on the fact that problem solutions must be adapted to specific problems, therefore modeling is often used to optimize production line operations, whether mathematical or computational [6][7].

Based on observation, the current KPI target for throughput rate is 36 ton/hour, but the actual achievement in the past 2 years is around 22 to 25 ton/hour. Through observations and interviews with seven relevant respondents, it was agreed that the production schedule affects the achievement of throughput levels. Therefore, this research seeks to develop the most appropriate model for estimating throughput rate achievement, so as to guide schedulers in making scheduling decisions based on the best expected results. Based by the fact that the production schedule's objective function which is the target and goal to be met, like makespan and inventory cost is a crucial component of scheduling [8]. The production system used in this research object is batch production make to order, whereas the effectiveness and efficiency this system are influenced by product transition [9] in handling variety of product, batch size, and complexity. Historical data is required to conduct a learning process in order to identify patterns and behaviors in the production system and to generate solutions. Machine learning can be employed to enhance transparency and predictive capabilities through applying sophisticated mathematical algorithms to identify cause-and-effect relationships and generate reliable predictions using data on hand [10][11]. In addition to achieving the research objectives of enhancing the efficiency of the production schedule, this study also seeks to offer an inside experience into the initial stages of the data mining process in the context of technological advancements.

The objective function that will be employed in this research is the prediction of production results, which will be used to evaluate the prediction of throughput rate achievement based on production schedule. Because improving the Production Planning and Control (PPC) function within a company can result in substantial overall improvements in the manufacturing system [12].The prediction will also be associated to optimal profit estimates, which will be determined by a variety of constraints, including production capacity and time. The throughput rate will be predicted by employing the machine learning process through data mining. The process will produce throughput rates that are driven by data, because data-driven production scheduling is more suitable for handling complicated production schedules that are impacted by customer objectives, constraints, and product customization [13][14]. Data mining is capable of facilitating the process of making decisions base knowledge in the scheduling process quickly by utilizing historical data [15]. Prediction patterns will be identified for each scheduled production schedule through the utilization of data mining in the form of the best prediction model among other models. Because this study will use several predictive models to predict throughput results, which will be selected based on the model performance result.

The novel aspect of this research is in the approach of predicting production throughput through finding prediction models that align with the system's capabilities. These models then paired with optimization models to maximize profits using operation research. Operations research is performed to achieve key objectives such as improving efficiency, reducing costs, and increasing effectiveness [16] through the production scheduling in this study. The production scheduling decision model should considers both maximizing throughput and estimates profit. This research contributes to the process of developing a production rate prediction model that can be implemented by considering optimal decisions in the animal feed industry. It's important to study how predictive analytics and optimization might be combined to maximize profits in animal feed production scheduling.

2. Method

The research method and the sequence of the research process are shown in Fig 1 The research began with a literature review and observation of the object of research based on field conditions and historical data.

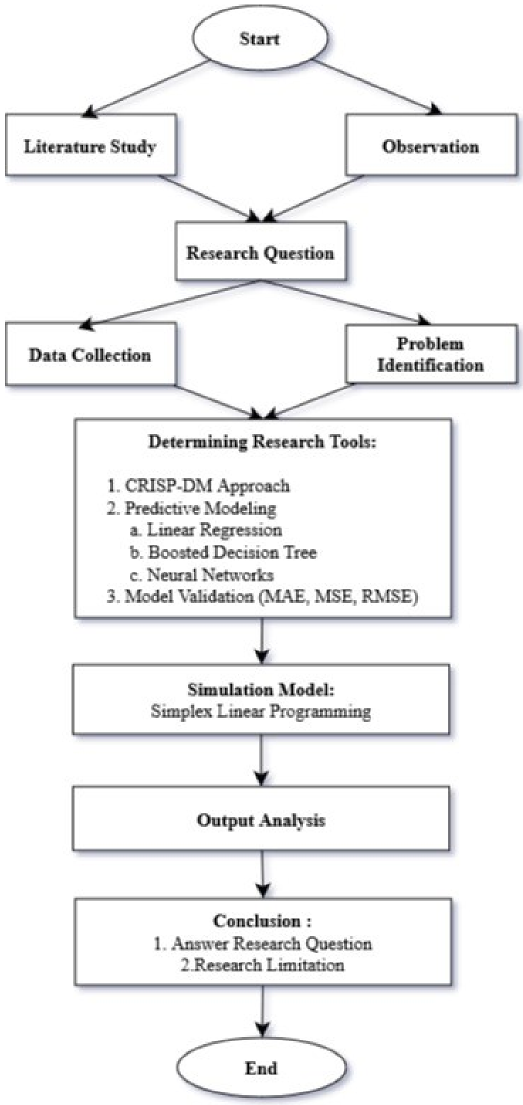


Fig 1. Research Flow Diagram

The next step is to determine the research questions to begin the process of data collection and problem identification. The research data collected is based on historical data in the form of a dataset size n = 680. The dataset contains information about the type of product that has been produced as an independent variable, and data about the production rate used as the dependent variable. This dataset has not gone through the data preparation and outlier elimination process, because the data preparation process is carried out at the CRISP-DM implementation stage. The CRISP-DM framework can help develop data-driven approaches for manufacturing systems which optimize operational performance and support industrial sustainability by being readily applicable and integrated into existing systems [17][18]. To further fulfill the objectives of this study, the following research questions will be addressed through data collection, data processing, and simulation.

Q1: How data mining supports the production system optimization process, in terms of scheduling.

Q2: How predictive analytics and optimization might be combined to maximize profits in animal feed production scheduling

10.12928/ijio.v6i1.11357

2.1. Data Collection and Problem Identification

Through study questions given to six respondents in different roles within the company, the variable of product type influencing throughput success has been validated. The six responders work in the departments of machine operators, production, and Production Planning and Inventory Control (PPIC). These variations in product types are critical to understanding the production dynamics within the feed mill. Each product form Pellet, Crumble, Fine, and Mash requires distinct processing steps, machinery settings, and cycle times, which directly impact the throughput and overall efficiency of the production line. The differences in processing requirements also require careful planning and optimization to ensure that the production schedules are aligned with the desired throughput rates, thereby minimizing bottlenecks and maximizing operational efficiency. By analyzing these product variations, the research aims to develop a more accurate predictive model for optimizing throughput in the animal feed industry.

The graph in Fig. 2 shows the throughput achievement over 10 months. It is clear that the target to be achieved is 36 tons per hour, while the actual performance is still far below the target. Due to the large gap between the target and the achievement, historical data analysis will be conducted to identify the factors contributing to this difference and assess the fairness of the acquisition process. In this analysis, the product type variable will be treated as the independent variable, and its relationship with the throughput rate, which is considered the dependent variable, will be explored. By applying linear regression analysis using SPSS, it was determined that 44.7% of the variability in the throughput rate can be explained by the type of product produced as carried out in Fitriah et.al research [19]. This shows that while the type of product has a significant impact on throughput, other factors also play a role, indicating the need for a more comprehensive analysis to identify additional variables that may affect production efficiency.

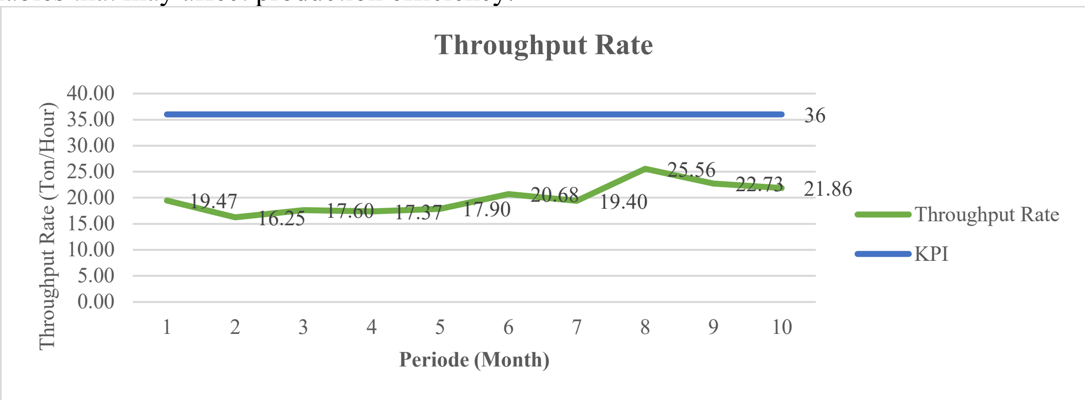


Fig 2. Production Throughput Gain at Research Object Company

2.2. Proposed Method

The proposed method involves utilizing the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to systematically analyze historical production data and develop predictive models [20][21][22] for production rate optimization in the animal feed industry. By applying data preprocessing, exploratory data analysis, and machine learning techniques within this framework, the method aims to describe the relationship of product variation in influencing production rate and integrate these insights into a dynamic production scheduling system. This approach will enable real-time adjustments to production plans, ensuring optimal production rates, improving overall efficiency, and directly involving profits in scheduling decisions.

In addition, the proposed method will provide a powerful decision support tool for schedulers, allowing them to anticipate potential congestion and adjust their delay anticipation measures accordingly. The inclusion of profit-driven criteria in the scheduling process also ensures that not only is the throughput rate optimized, but the overall profitability of the production line is enhanced, aligning operational efficiency with business goals. This method builds a system that can adapt to modifications in the production schedule, enabling it to offer the best recommendations both periodically and in real-time.

2.3. CRISP-DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology in Fig 3, which is widely used in the social sciences and was developed in the mid-1990s [23].

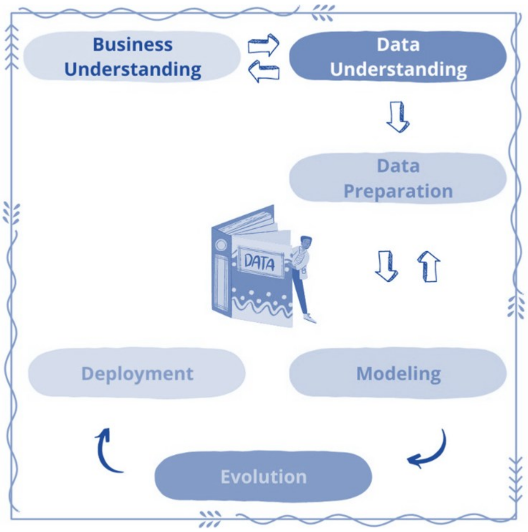


Fig 3. CRISP-DM Methodology

The CRISP-DM method consists of six main stages that form a systematic framework for the data mining process. These stages include [23]:

1. Business Understanding: Understanding the business objectives and needs to determine how data mining can provide suitable solutions.
2. Data Understanding: Collecting and exploring initial data to gain a deep understanding of the data and the problems to be addressed.
3. Data Preparation: Preparing data for analysis, including cleaning, transforming, and integrating data to make it ready for modeling.
4. Modeling: Applying data mining techniques and algorithms to build predictive or classification models in line with business objectives.
5. Evaluation: Assessing the performance of the developed model to ensure it meets business objectives and provides reliable results.
6. Deployment: Implementing the evaluated model into a production environment, and monitoring and updating the model as needed based on changing requirements or data conditions.
7. Each stage is interrelated and functions to ensure that the data mining process is thorough and effective, from initial understanding to practical implementation.

2.4. Predictive Models

Some of the predictive models that will be used for the machine learning process are as follows.

1. Linear Regression
   Regression is a statistical method used to build and obtain a mathematical equation model that aims to predict or estimate the value of Y (one or more targets) based on input variables X that have certain dimensions and are continuous[24]. This process involves to understand how strongly the independent variables influence or explain variations in the dependent variable. The resulting regression model allows for more precise decision making based on existing data, and can be used to estimate future values, understand patterns of relationships between variables, and identify factors that have a significant impact on the variables being studied.
2. Boosted Decision Tree

10.12928/ijio.v6i1.11357

Decision tree algorithms and boosting techniques are combined to create boosted regression tree (BRT) models. In order to increase the model's accuracy, BRTs fit numerous decision trees repeatedly [25]. Boosted tree diagram depicts a model that consists of multiple decision trees combined to enhance the overall model performance. In a boosted tree diagram, each new decision tree is added to correct errors made by the previous trees. This diagram shows how each individual tree contributes to the final model and how the weighting of these trees is adjusted during the boosting process

3. Neural Networks

   Neural networks (NNs) are a method of choice for developing learning algorithms due to their ability to handle complex tasks and data[26]. Neural networks are designed to estimate unknown functions based on observed data to produce accurate predictions.

2.5. Model Validation

The model validation process is carried out in 2 stages, the first is validating the prediction model, the second is simulating the model to prove that the prediction model is valid and can be implemented.

2.5.1 Predictive Model Validation

Calculating Mean Absolter Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) is essential for assessing predictive model performance [27][5][28]. Quantifying the differences between expected and actual values helps these measures assess model accuracy and reliability. These computations are critical for measuring model performance and guiding modifications since MAE represents average absolute differences, MSE measures average squared errors, and RMSE measures error in the same units as the original data.

$$Mean\ Absolute\ Eror\ (MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$Mean\ Squared\ Eror\ (MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$Root\ Mean\ Squared\ Eror\ (RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where n is number of data points; y_i is actual values; $\hat{y}_i$ is predicted values; $|y_i - \hat{y}_i|$ is absolute difference between actual and predicted values; $(y_i - \hat{y}_i)^2$ is squared difference between actual and predicted values.

2.5.2 Model Simulation

The simulation method is facilitated by the Solver add-in in Microsoft Excel [29][30], allowing for simulations aimed at maximizing throughput rate and profit. The constraints applied include a maximum capacity of 30 tons per hour and a working schedule of 9 hours per day. Additionally, objective function introduced with specifying the maximum profit for each product being produced. This setup enables the identification of optimal production strategies that balance throughput and profitability within the defined operational limits.

$$Max\ (Z) = C_1X_1 + C_2X_2 + \ldots + C_nX_n \quad (4)$$

3. Results and Discussion

3.1. CRISP-DM Implementation

According to the Cross Industry Standard procedure for Data Mining (CRISP-DM) methodology, the entire data mining procedure can be arranged into six stages.

A. Business Understanding

Conventional production scheduling plans adjust production lines based on historical experience, but production conditions will differ significantly from predicted results which will affect schedules and delivery times [31]. To address this issue, more dynamic and data-driven methods are needed, such as the application of machine learning technology or real-time data analysis, which can provide more accurate predictions and faster responses to changes in production conditions. With the solving process sequence shown in Fig 4.

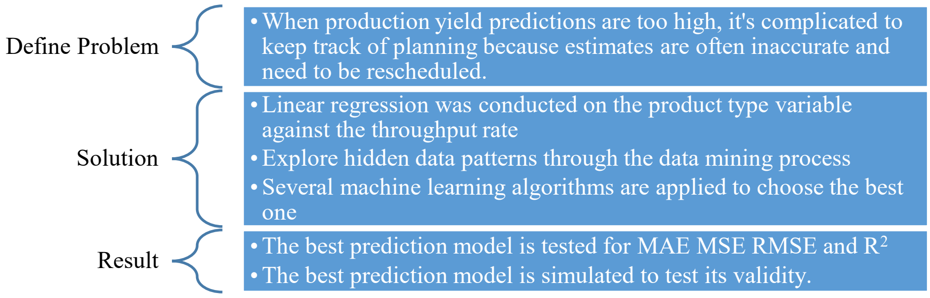


Fig 4. Business Understanding Flow

B. Data Understanding

By applying linear regression analysis using SPSS, it was determined that 44.7% of the variability in the throughput rate can be explained by the type of product produced. This finding underscores the significant influence that product type has on the throughput rate, suggesting that different products inherently require varying levels of time, resources, and processes, which directly impacts production efficiency. However, the fact that over half of the variability (55.3%) remains unexplained highlights the presence of other influential factors. These could include elements such as equipment efficiency and workforce skill levels.

C. Data Preparation

The dataset preparation process involves several important steps to ensure that the data is ready for further analysis. Starting with the initial data of 680 data, outlier identification was carried out using a boxplot diagram in Fig 5 to remove data that was not appropriate or extreme.

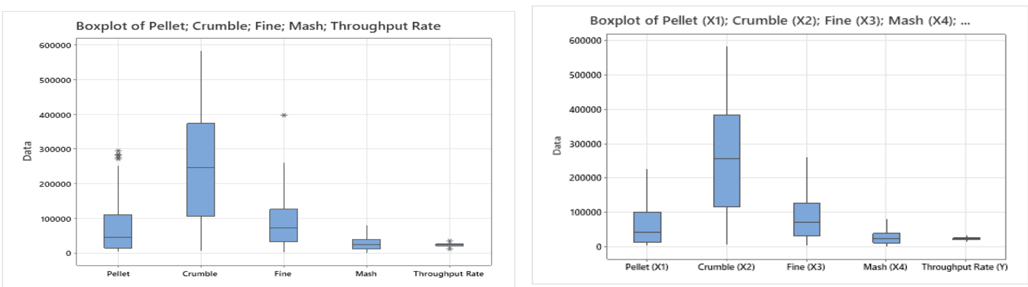
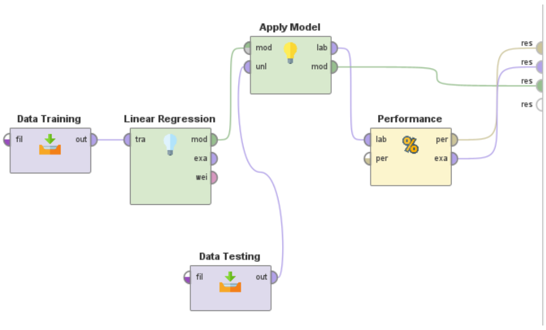


Fig 5. Datasets Before (left) and After (right) Eliminate Outlier

In this process, 571 data were detected as outliers, which were then removed from the dataset. After this cleaning process, 109 data remained which were considered clean and suitable for further analysis. Eliminating outliers plays a critical role in increasing the prediction model's accuracy, reducing noise in the data, normalizing the dataset, focused on more representative data, minimizing the possibility of overfitting, and enhancing managerial decision-making [32][33]. A boxplot diagram was used to test the dataset in this study, and the results of the normalization test were obtained with P-value less than 0.05

D. Modeling

The RapidMiner program is being used by multiple operators to construct prediction models. The first operator is the data access operator, which has two components: a training data input and an additional testing data component. Then comes the modeling operator, which links training data, algorithms, and testing data using the Apply Model operator. This operator includes linear regression in Model 1 in Fig 6, gradient boosted tree in Model 2 in Fig 6, and neural network models in Model 3 in Fig 8. The performance operator is the last step in testing the prediction model and determining the R2, MAE, MSE, and MRSE values.



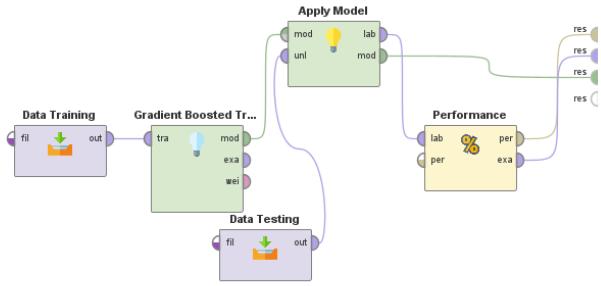


Fig 6. Data Mining Model 1- Linear Regression **Fig 7.** Data Mining Model 2-Gradient Boosted Tree

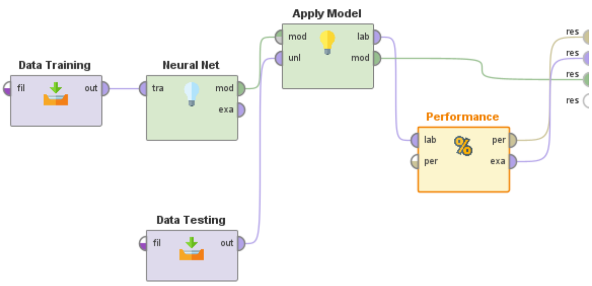


Fig 8. Data Mining Model 3- Neural Networks

E. Evaluation

The selection of the best model is determined through the evaluation of model performance using metrics such as MAE, MSE, and RMSE shown in the Table 1. By comparing these metrics, we identify the model that exhibits the lowest error rates and best balances accuracy and reliability, ensuring the most effective predictive performance. Model 1 was chosen despite not having the lowest MAE, RMSE, or MSE. The high R^2 value of the model indicates that it provides a better overall fit to the data, explaining a higher proportion of the variability compared to the other models. Therefore, Model 1 was chosen due to its balanced performance across metrics and superior explanatory power.

Table 1. Predictive modeling error test

Models	MAE	MRSE	MSE	R^2
Model 1	5.186,46	5.970,32	1,59	0,72
Model 2	4.172,68	4.849,83	1,21	0,07
Model 3	6.111,98	6.501,97	1,87	0,00

Model 1 has an MAE of 5,186.46, which is higher than Model 2 (4,172.68) but lower than Model 3 (6,111.98). A lower MAE indicates fewer errors on average, but in this case, Model 1's MAE is relatively balanced compared to the others. Model 1's MSE is 1.59, compared to Model 2's 1.21 and Model 3's 1.87. Although Model 2 has a slightly lower MSE, the difference is not substantial enough to outweigh other factors. Model 1 has an RMSE of 5,970.32, which is also higher than Model 2 (4,849.83) but lower than Model 3 (6,501.97). RMSE penalizes larger errors more heavily, so a lower RMSE is preferred. Despite having a higher RMSE than Model 2, Model 1 still performs better than Model 3. odel 1 has an R^2 of 0.72, indicating that it explains 72% of the variability in the data. This is significantly higher than Model 2 (0.07) and Model 3 (0.00), meaning Model 1 provides a better fit for the data and accounts for more variance. Model 1 was chosen because it has the best test error results. Model 1 is also suitable to be implemented as a production rate prediction model because the model performance is simple and easy to use. However, linear regression methods are very sensitive to data distribution and require linearity checks in the regression task. As a side effect, the data transformation requirements are more stringent, often requiring multiple tests to find the appropriate model transformation [34]. Therefore, in the simulation stage, it will be tested with the simplex optimization method by providing constraints on the system.

F. Deployment

The selected model produces an objective function to maximize the throughput rate. In addition to producing an objective function, it also produces a predicted value that appears when the independent and dependent variable data are integrated into the RapidMiner output in Fig 9.

$$y = 0{,}026\ X1 + 0{,}016\ X2 + 0{,}010\ X3 + 16.795{,}657$$

where y is throughput rate prediction (ton/hour), $X1$ is target production quantity of pellet feed (kg), $X2$ is target production quantity of crumble feed (kg), $X3$ is target production quantity of fine feed (kg).

Ro... ↑	Pellet	Crumble	Fine	Mash	prediction(Throughput Rate (Y))	Throughput Rate (Y)
1	85400	54950	106800	52200	20944.711	33261
2	9700	118800	71700	20100	19610.181	24478
3	43600	43950	19800	56350	18818.989	18189
4	94450	32550	90400	17950	20670.669	26150
5	84950	96500	34550	20050	20867.615	26228
6	111450	10900	90600	24100	20780.518	26339
7	124600	107800	12600	16050	21866.160	29006
8	47600	95800	74450	4050	20271.831	24656
9	139400	9300	76050	4000	21344.358	25417
10	17300	70550	11900	47950	18466.613	16411

Fig 9. Output of throughput rate prediction on RapidMiner software

The residual histogram in Fig 10 is used to analyze the gap between the predicted and actual results, specifically to assess whether the residuals are normally distributed. This histogram graph displays the distribution of residuals, helping to identify any deviations from normality. Although the data appears to be generally in line with a normal distribution, it is important to note that the zero point does not necessarily align with the peak of the curve. This misalignment may indicate some level of bias in the model's predictions or suggest areas where the model could be improved.

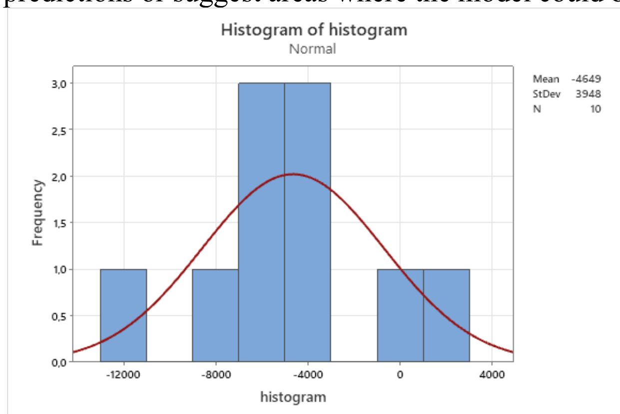


Fig 10. Prediction residue histogram

10.12928/ijio.v6i1.11357

3.2. Simulation

There are three simulations conducted in this analysis. The first simulation involves optimization while considering sales requirements, aiming to align production with demand. The second simulation examines the scenario where sales requirements are disregarded, providing insight into the model's performance without the constraint of meeting specific sales targets. The final simulation focuses on maximizing profit by using a profit optimization objective function. This approach evaluates the best possible outcome in terms of financial gain, without necessarily considering sales constraints or demand. Each simulation offers a different perspective on how various factors influence the overall performance and efficiency of the system. An optimization method called linear programming is used to determine a linear objective function's ideal value when faced with limitations [35] [36].

The conclusion of the first simulation in Table 2 indicates that optimal throughput is achieved by focusing exclusively on the production of crumble and fine feed as needed, with the remainder of the production dedicated to pellet feed. This strategy is based on the observation that pellet feed has the highest coefficient value, making it the most economically advantageous choice. By prioritizing the production of pellet feed, the system maximizes throughput efficiency while effectively meeting the demand for crumble and fine feed. This approach not only enhances overall production performance but also aligns with the objective of optimizing resource allocation and maximizing output based on the relative value of each feed type.

Table 2. First simplex simulation model

Throughput Rate Prediction Objective Function (y) :

$$0{,}026\ X1 + 0{,}016\ X2 + 0{,}010\ X3 + 16.795{,}657$$

Objective Function :	22.220	Tons/hour
Time Constraint :	9	Hours/Day
Capacity Constraint :	30.000	Tons/hour

Product	Planning Production Quantity	Products for Sale	Profit Estimation	
	kg	kg	Unit Profit (Rp)	Total Profits (Rp)
X1:	150.000	40.000	500	75.000.000
X2 :	60.000	60.000	650	39.000.000
X3 :	60.000	60.000	750	45.000.000
Total:	**270.000**	**270.000**		**159.000.000**

In the second simulation in Table 3**Error! Reference source not found.**, the sales requirements constraint was not applied, revealing that optimal results can be achieved when the production system focuses exclusively on producing a single type of product, specifically pellet feed. Without the constraint of meeting specific sales demands, the system demonstrates that concentrating resources on the production of pellet feed alone leads to the highest efficiency and throughput. This result highlights the potential benefits of streamlining production processes and focusing on the most profitable or highest-value product, thereby optimizing overall performance and output in the absence of sales constraints.

Table 3. Second simplex simulation model

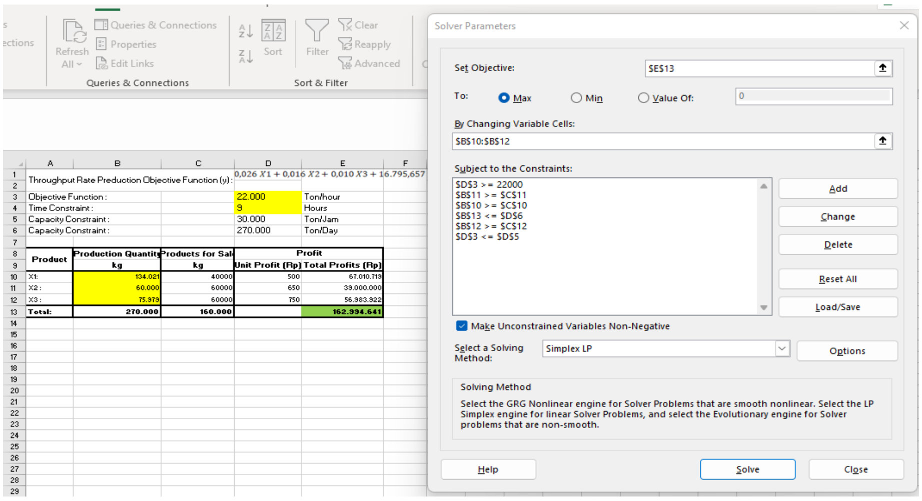
Throughput Rate Prediction Objective Function (y) :	$0{,}026\ X1 + 0{,}016\ X2 + 0{,}010\ X3 + 16.795{,}657$	
Objective Function :	23.780	Tons/hour
Time Constraint :	9	Hours/day

Capacity Constraint :		30.000		Tons/Hour
Product	**Planning Production Quantity**	**Products for Sale**	**Profit-Estimation**	
	kg	**kg**	**Unit Profit (Rp)**	**Total Profits (Rp)**
X1:	270.000	0	500	135.000.000
X2 :	0	0	650	0
X3 :	0	0	750	0
Total:	**270.000**	**270.000**		**135.000.000**

The final simulation in .

Fig *11.* Simulation Modeling Using Solver Ms. Excel

Table *4* involves using a profit maximization objective function. This third simulation yields particularly compelling results as it projects the optimal profit achievable under the given conditions. By focusing on maximizing profit, the simulation demonstrates how adjusting production strategies to prioritize financial gain can lead to significant improvements in overall profitability. The results



underscore the effectiveness of profit-centered optimization in identifying the most advantageous production scenario, providing valuable insights into how to achieve the highest possible return on investment while balancing operational constraints and opportunities. The results of the third simulation modeling using excel solver are shown in the Fig 11.

Fig 11. Simulation Modeling Using Solver Ms. Excel

Table 4. Third simplex simulation model

Throughput Rate Preduction Objective Function (y) :		$0{,}026\ X1 + 0{,}016\ X2 + 0{,}010\ X3 + 16.795{,}657$		
Objective Function :		22.000		Tons/hour
Time Constraint :		9		Hours/Day
Capacity Constraint :		30.000		Tons/Hour
Capacity Constraint :		270.000		Tons/Day
Product	**Planning Production Quantity**	**Products for Sale**	**Profit Estimation**	

 10.12928/ijio.v6i1.11357

	kg	kg	Unit Profit (Rp)	Total Profits (Rp)
X1:	134.021	40000	500	67.010.719
X2 :	60.000	60000	650	39.000.000
X3 :	75.979	60000	750	56.983.922
Total:	**270.000**	**160.000**		**162.994.641**

3.3. Discussion

The purpose of this research has similarities with the research of Zhao et. al. [15], which uses the data mining process to reduce makespan in a job shop scheduling system. The research aims to minimize makespan by combining the process of data mining, discrete event simulation, and dispatching rules (DRs). While this study aims to develop a throughput maximization model, by combining data mining and simplex linear programming on a serial batch production system. However, both studies have concluded that data mining activities can help in scheduling activities and extracting knowledge based on the availability of data and appropriate analysis methods.

This research method is different from that conducted by Hiller et.al. [28], where the research predicts the need for process time at each stage in fulfilling customer orders starting from receiving orders to completing order completion. Hiller et.al [28] research also used machine learning in conducting the research, with the aim of being able to provide time predictions to customers for orders. While the current research focuses on improving the accuracy of production scheduling that is longer than the daily schedule such as the weekly plan, detecting problems in the production system that cannot be detected early, and updating knowledge in the organization. In applying the predict the optimize approach, research by Tian et al [34] performs data transformation techniques as an optimization technique to overcome values that are zero and potentially reduce the reliability of the prediction model. However, in this study, to meet the optimal criteria, the selected prediction model, namely the linear regression model, is tested again with a simplex approach to provide constraints that affect the optimal production plan decision. The constraints are the time and machine capacity of production, maximum profit objective function considerations, and sales requirements that must be fulfilled. Sangngern and Boonperm [37] in optimization using simplex, requires several iterations of the initial basis in the objective function of the m{1odel to then do many iterations until it approaches the optimal solution. While this research does not focus on model iteration, because the objective function as an initial basis has been determined through machine learning. This research produces a simpler optimization process by using a combination of machine learning and linear programming or simplex method.

Several studies have defined makespan as the objective function to be minimized through the machine learning process [38][39][40]. Makespan prediction usually aims to minimize the overall production duration, while throughput rate prediction aims to maximize the output per unit of time. In this case study, the duration of production duration is not a variable that can be adjusted, but for animal feed production, the measures of productivity are seen from the throughput rate. The limitation of this research is that the initial base is determined through machine learning without prioritizing the iteration process. This may cause limitations if the initial base is not fully optimized or if further iteration processes may provide better solutions. However, for studies and implementations with less risk of prediction errors, the method in this research is still suitable for similar optimization cases such as similar manufacturing systems with limited supporting historical data.

4. Conclusion

After going through 6 stages in the data mining process with the CRISP-DM method, it was found that linear regression modeling using independent variables of product type and dependent variables of throughput rate is valid for use in predicting throughput rate. The independent variables are used in the study because they have an R test which shows that 44.7% of the variability in the throughput rate can be explained by the type of product produced. After going through the data preparation process and eliminating outliers, the dataset consisting of 109 data with 10% being training data, is able to generate prediction output based on 3 different models. Then the prediction model testing is carried out by comparing the values of MAE, MSE, RMSE, and R^2, and shows that the linear regression model has the best prediction results and is balanced between its accuracy and variability.

Throughput level prediction can be developed into profit prediction by developing an objective function in the implementation of simplex linear programming as an operation research process. This combination can be used as an additional consideration for schedulers in determining production scheduling decisions. The results of this study showed that to predict a more optimal production schedule performance, several constraints such as production capacity, time capacity, and additional information such as cost/profit are required. Accordingly, the novelty of this research is how the results of machine learning can be simulated to operation research.

References

[1] A. Nubahriati, D. Natalia, and A. Sutomo, "Manufacturing Cycle Effectiveness dalam Meningkatkan Kinerja Studi Kasus Pada Bengkel Sinar Las Di Kota Watampone," *Journal of Applied Management and Business Research*, vol. 2, no. 1, pp. 83–90, 2022, doi: 10.38531/jambir.v2i1.48.

[2] M. Gopalakrishnan, M. Subramaniyan, and A. Skoogh, "The Management of Operations Data-driven machine criticality assessment – maintenance decision support for increased productivity," *Production Planning & Control*, vol. 33, no. 1, pp. 1–19, 2022, doi: 10.1080/09537287.2020.1817601.

[3] M. Subramaniyan, A. Skoogh, J. Bokrantz, M. A. Sheikh, M. Thürer, and Q. Chang, "Artificial intelligence for throughput bottleneck analysis – State-of-the-art and future directions," *J Manuf Syst*, vol. 60, no. August, pp. 734–751, 2021, doi: 10.1016/j.jmsy.2021.07.021.

[4] W. Chen, H. Liu, and E. Qi, "Discrete event-driven model predictive control for real-time work-in-process optimization in serial production systems," *Journal of Manufacturing Systems*, vol. 55, no. June 2019, pp. 132–142, 2020, doi: 10.1016/j.jmsy.2020.03.002.

[5] T. Hiller, L. Deipenwisch, and P. Nyhuis, "Systemising Data-driven Methods for Predicting Throughput Time within Production Planning & Control," *IEEE International Conference on Industrial Engineering and Engineering Management*, vol. 2022-Decem, pp. 716–721, 2022, doi: 10.1109/IEEM55944.2022.9989885.

[6] Z. Kang, C. Catal, and B. Tekinerdogan, "Machine learning applications in production lines: A systematic literature review," *Comput Ind Eng*, vol. 149, no. August, p. 106773, 2020, doi: 10.1016/j.cie.2020.106773.

[7] H. Sun, "Optimizing manufacturing scheduling with genetic algorithm and LSTM neural networks," *International Journal of Simulation Modelling*, vol. 22, no. 3, pp. 508-519, 2023, doi: 10.2507/IJSIMM22-3-CO13.

[8] A. S. Afolalu, O. M. Ikumapayi, and S. Ongbali, "Analysis of Production Scheduling Initiatives in the Manufacturing Systems," *International Journal of Mechanical and Production*, vol. 10, no. 3, pp. 1301–1313, 2020.

[9] Zainab Eldardiry, "A Conceptual Framework for Reducing Changeover Time in Batch Production Facilities," *International Journal of Engineering Research and*, vol. V10, no. 01, pp. 236–240, 2021, doi: 10.17577/ijertv10is010102.

[10] G. Schuh, G. Lukas, J. Schweins, J. Trisjono, and J. Frank, “Calculation of product service systems in single and small batch production,” *Journal of Revenue and Pricing Management*, no. 0123456789, 2024, doi: 10.1057/s41272-023-00455-5.

[11] C. Gröger, F. Niedermann, and B. Mitschang, "Data mining-driven manufacturing process optimization," in *Proceedings of the World Congress on Engineering*, vol. 3, pp. 4-6, July 2012, available: https://www.iaeng.org/.

[12] J. P. Usuga Cadavid, S. Lamouri, B. Grabot, R. Pellerin, and A. Fortin, “Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0,” *J Intell Manuf*, vol. 31, no. 6, pp. 1531–1558, 2020, doi: 10.1007/s10845-019-01531-7.

[13] Y. Ma, S. Li, F. Qiao, X. Lu, and J. Liu, “A data-driven scheduling knowledge management method for smart shop floor,” *Int J Comput Integr Manuf*, vol. 35, no. 7, pp. 780–793, 2022, doi: 10.1080/0951192X.2022.2025622.

[14] H. Kim, D. E. Lim, and S. Lee, “Deep Learning-Based Dynamic Scheduling for Semiconductor Manufacturing with High Uncertainty of Automated Material Handling System Capability,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 1, pp. 13–22, 2020, doi: 10.1109/TSM.2020.2965293.

[15] A. Zhao, P. Liu, X. Gao, G. Huang, X. Yang, Y. Ma, and Y. Li, "Data-mining-based real-time optimization of the job shop scheduling problem," Mathematics, vol. 10, no. 23, p. 4608, 2022, doi: 10.3390/math10234608.

[16] S. K. Tripathi and R. Kumar, “A Short Literature on Linear Programming Problem,” *EAI Endorsed Transactions on Energy Web*, vol. 10, pp. 1–5, 2023, doi: 10.4108/ew.4516.

[17] H. Jacobsen and K. H. Tan, “Improving food safety through data pattern discovery in a sensor-based monitoring system,” *Production Planning and Control*, vol. 33, no. 16, pp. 1548–1558, 2022, doi: 10.1080/09537287.2021.1882691.

[18] O. J. Fisher *et al.*, “Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems,” *Computers & Chemical Engineering*, vol. 140, 2020, doi: 10.1016/j.compchemeng.2020.106881.

[19] M. Faishal, et al., "Investigating factors of the purchase intention of slaughterhouses for Halal Certification in Yogyakarta, Indonesia," *Multidisciplinary Science Journal*, vol. 6, no. 12, p. 2024266, 2024, doi: 10.31893/multiscience.2024266.

[20] R. Wirth and J. Hipp, “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959, pp. 29–39, 2000, [Online]. Available: https://www.researchgate.net/.

[21] J. Wegener, S. Vanputten, J. Neubeck, and A. Wagner, “Data Mining as an Enabler for Customer Data Driven Vehicle Development,” *Tongji Daxue Xuebao/Journal of Tongji University*, vol. 49, pp. 1–10, 2021.

[22] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, “DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model,” *Procedia CIRP*, vol. 79, pp. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.

[23] E. Kesriklioğlu, E. Oktay, and A. Karaaslan, “Predicting total household energy expenditures using ensemble learning methods,” *Energy*, vol. 276, no. March, 2023, doi: 10.1016/j.energy.2023.127581.

[24] C. K. Eng and H. M. Asih, "An Integrated Robust Optimization model of capacity planning under demand uncertainty in electronic industry," *International Journal of Mechanical & Mechatronics Engineering*, vol. 15, no. 3, pp. 88-96, 2015, Available: https://www.researchgate.net/.

[25] E. Jumin, F. B. Basaruddin, Y. B. M. Yusoff, S. D. Latif, and A. N. Ahmed, “Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia,” *Environmental Science and Pollution Research*, vol. 28, no. 21, pp. 26571–26583, 2021, doi: 10.1007/s11356-021-12435-6.

[26] R. Devore, B. Hanin, and G. Petrova, “Neural network approximation,” *Acta Numerica*, vol. 30, pp. 327–444, 2021, doi: 10.1017/S0962492921000052.

[27] J. J. Jui, M. M. Imran Molla, B. S. Bari, M. Rashid, and M. J. Hasan, “Flat Price Prediction Using Linear and Random Forest Regression Based on Machine Learning Techniques,” *Lecture Notes in Electrical Engineering*, vol. 678, no. March 2021, pp. 205–217, 2020, doi: 10.1007/978-981-15-6025-5_19.

[28] T. Hiller, T. M. Demke, and P. Nyhuis, “Throughput Time Predictions Along the Order Fulfilment Process,” *IEEE Access*, vol. 12, pp. 9705–9718, 2024, doi: 10.1109/ACCESS.2024.3353029.

[29] L. Briones, V. Morales, J. Iglesias, G. Morales, and J. M. Escola, “Application of the microsoft excel solver tool in the optimization of distillation sequences problems,” *Computer Applications in Engineering Education*, vol. 28, no. 2, pp. 304–313, 2020, doi: 10.1002/cae.22193.

[30] R. A. C. Leuveano, H. M. Asih, M. I. Ridho, and D. A. Darmawan, "Balancing inventory management: Genetic algorithm optimization for a novel dynamic lot sizing model in perishable product manufacturing," *Journal of Robotics and Control (JRC)*, vol. 4, no. 6, pp. 878-895, 2023, doi: 10.18196/jrc.v4i6.20667.

[31] C. Wang and S. Zhou, “Approximate multivariate distribution of key performance indicators through ordered block model and pair-copula construction,” *IISE Trans*, vol. 51, no. 11, pp. 1265–1278, 2019, doi: 10.1080/24725854.2018.1550826.

[32] S. Moon and J. Lee, “Research on Improving Fetal Health Prediction Model Using Optimal Fetal Feature Selection Technique,” *Journal of Korean Institute of Communications and Information Sciences*, vol. 49, no. 2, pp. 270–276, 2024, doi: 10.7840/kics.2024.49.2.270.

[33] J. E. Fontecha, P. Agarwal, M. N. Torres, S. Mukherjee, J. L. Walteros, and J. P. Rodríguez, “A Two-Stage Data-Driven Spatiotemporal Analysis to Predict Failure Risk of Urban Sewer Systems Leveraging Machine Learning Algorithms,” *Risk Analysis*, vol. 41, no. 12, pp. 2356–2391, 2021, doi: 10.1111/risa.13742.

[34] X. Tian, Y. Guan, and S. Wang, “Data Transformation in the Predict-Then-Optimize Framework: Enhancing Decision Making under Uncertainty,” *Mathematics*, vol. 11, no. 17, 2023, doi: 10.3390/math11173782.

[35] H. M. Asih and C. K. Eng, "Cost-volume-profit analysis for uncertain capacity planning: A case study paper," in *Proceedings of the Second Asia Pacific International Conference on Industrial Engineering and Operations Management, Surakarta, Indonesia*, 2021, available: http://ieomsociety.org/.

[36] A. Gleixner and D. E. Steffy, “Linear programming using limited-precision oracles,” *Math Program*, vol. 183, no. 1–2, pp. 525–554, 2020, doi: 10.1007/s10107-019-01444-6.

[37] K. Sangngern and A. A. Boonperm, “A new initial basis for the simplex method combined with the nonfeasible basis method,” In *Journal of Physics: Conference Series*, vol. 1593, no. 1, 2020, doi: 10.1088/1742-6596/1593/1/012002.

[38] H. Yamashiro and H. Nonaka, “Estimation of processing time using machine learning and real factory data for optimization of parallel machine scheduling problem,” *Operations Research Perspectives*, vol. 8, no. July, p. 100196, 2021, doi: 10.1016/j.orp.2021.100196.

[39] D. Tremblet, S. Thevenin, and A. Dolgui, “Makespan estimation in a flexible job-shop scheduling environment using machine learning,” *International Journal of Production Research*, vol. 62, no. 10, pp. 3654–3670, 2024, doi: 10.1080/00207543.2023.2245918.

[40] A. Oktafiani and M. N. Ardiansyah, “Scheduling Splitable Jobs on Identical Parallel Machines to Minimize Makespan using Mixed Integer Linear Programming,” *International Journal of Innovation in Enterprise System*, vol. 7, no. 01, pp. 41–54, 2023, doi: 10.25124/ijies.v7i01.190.