Data Mining Using Predictive Analysis for Raw Material Stock in PT. PKT: An Application of The CRISP-DM Methodology

Andrean Wijaya[1], Fakhrur Razi[2], Kasthalani[3]

[1,2,3]Bina Nusantara University

Email: andreanw96@binus.ac.id[1], fakhrur.razi@binus.ac.id[2], kasthalani@binus.ac.id[3]

Abstrak: Tujuan dari penulisan ini adalah untuk mengevaluasi beberapa model pembelajaran mesin dengan metodologi CRISP-DM guna menentukan, model terbaik untuk memprediksi status stok tingkat persediaan di masa depan sehingga ketersediaan bahan baku selalu mampu memenuhi permintaan produksi dan penjualan. . Penelitian ini akan dianalisis menggunakan Analisis Deskriptif dan Prediktif menggunakan teknik data mining dengan Metodologi CRISP-DM. Operator pertama yang digunakan dalam model ini adalah Generalized linear model (GLMs), Algoritma ini menyesuaikan model linear umum ke data dengan memaksimalkan kemungkinan log yang merupakan perluasan dari model linear tradisional. Model ini memiliki root mean squared error sebesar 0,279 dan akurasi prediksi model ini sebesar 91,11%. Dengan analisis tersebut, diasumsikan PKT dapat memprediksi perkiraan jumlah pembelian bahan baku, target produksi dan perkiraan potensi penjualan di masa yang akan datang.
Kata Kunci: Metodologi CRISP-DM, Pupuk, Inventaris, *Machine Learning*, *Generalized Linear Model.*

Abstract: *The purpose of this paper is to evaluate several machine learning models under the CRISP-DM methodology in order to determine, the best model for predicting inventory level stock status in the future so availability of raw material always able to fulfil production and sales demand. This research will be analyzed using Descriptive and Predictive Analysis using data mining techniques with the CRISP-DM Methodology. The first operator used in this model is Generalized linear models (GLMs), This algorithm fits generalized linear models to the data by maximizing the log-likelihood an extension of traditional linear models. This model has 0.279 root mean squared error and prediction accuracy of this model is 91,11%. With this analysis, assuming PKT can predict the estimated number of raw material purchases, production target and potential sales forecast in future.*
Keywords: *CRISP-DM Methodology, Fertilizer, Inventory, Machine Learning, Generalized Linear Models.*

INTRODUCTION

PT. PKT is one of strategic industrial company which is subsidiary of PT. PI as holding company who become the biggest organic fertilizer producer in Indonesia. The company was started after the success of floating fertilizer factory built by one of the state-owned enterprises in 1975. Then the government which represented by the Ministry of Industry took over the factory and relocated it to the land. PKT was officially established on December 7, 1977.

Currently, PKT has 13 plants spread around Indonesia including 5 Ammoniac plants with capacity 2,74 million ton per year which is the biggest in South-East Asia, 5 urea plants with capacity 3,43 million ton per year and 3 NPK plants with total capacity 300 thousand-ton per year.

Uncertainty of raw material requirement with several factors as considerations such as production plan, warehouse capacity, lifetime material and its fluctuating price. Domestic demand of fertilizer is significantly volatile following government regulation. It is also affected by the growing season and weather season. It needs cross-function collaboration involving multiple divisions of the company to make a proper plan with commitment to support each other. Continuous evaluation and improvement as part of tactical plan to be aligned with company strategy.

This research will be analyzed using Predictive Analysis. It uses historical and current sets of data of inventory stock, purchasing and sales to define how to predict inventory level stock status in the future so availability of raw material always able to fulfill production and sales demand. With this analysis, assuming PKT can predict when and estimated number of raw material purchases. Also, PKT can use this analysis to align supply of raw material with demand of production to reduce potential loss due to dead stock and over buying.

LITERATURE REVIEW

a. Data Science and Data Mining

Processing massive amounts of data to support business analysis and decision process has become a priority of IT (Information Technology) in every industry. We call it Data Science that is discipline to gain valuable insight from data by mathematical and analytical model and application (Schröer et al., 2021). One of the data science processes is Data mining, a creative process which combination of a few different skills and knowledge. At this moment, no standard framework carry out data mining projects that the success or failure of a data mining project would depends on person. Data mining needs a standard approach which will help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience(Huber et al., 2019)

Data Mining finds the pattern and information from selected data set using technique and specific method. Data mining is divided to 5 phases that is estimation, prediction, classification, clustering, and association. The most used data mining technique is classification which is used to map the data into groups and defined classes (Hasanah et al., 2021)

b. Cross-Industry Standard Process for Data Mining

The CRISP-DM (Cross Industry Standard Process for Data Mining) project addressed parts of these problems by defining a process model which provides a framework for carrying out data mining projects which is independent of both the industry sector and the technology used. With CRISP-DM process model, data mining process can be less costly but more reliable, repeatable, manageable, and definitely faster (Wirth & Hipp, n.d.)

CRISP-DM is an industry-independent process model for data mining which consisted of six iterative phases start from business understanding until deployment. Below table describes the main idea, tasks, and output of these phases shortly, based on the user guide of CRISP-DM (Schröer et al., 2021)

Phase	Main Idea
Business Understanding	How to understand the objectives and requirements. Any good project starts with a deep understanding of the customer's needs.
Data Understanding	Collect related data from data sources, assessment, describing, and verify data quality.
Data Preparation	Data selection to define inclusion and exclusion criteria.
Data Modelling	Iterate model building and assessment to find the best models. This process can take more time to do iterating until it comes with "good enough" model, then there is further improve the model in next iterations.
Evaluation	Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed

Deployment	A model is not particularly useful unless the customer can access its results

CRISP-DM portrays the project as a cyclical process, in which all phases are tuned towards achieving the core business goal. For most businesses, structuring and prioritizing their business project gets tedious. By using CRISP-DM as business analysis strategy, a project follows a clear roadmap which helps in achieving desired goal (Parate, n.d.)

c. Supply Chain Management

In this globalization era, industry need to adopt a systemic concept of supply chain strategy management. It is needed to anticipate reduction of entry barriers into industries, technological advancement, increased information/knowledge transfer, and emerging markets rejuvenating mature products and industries.

Fundamental of supply chain management is developed based on philosophy of business organization and the idea of partnership between marketing channel linkage to entities. If refers to traditional models of business organization, the basic concept of this are by maximizing their revenues and minimizing their costs. To achieve this goal, there might be another entity that received the withdraws and take disadvantage of this. And now, global supply chain management model rephrases the goal is to maximize profit by enhanced the competitiveness that who can achieve lower cost to serve and shortest time-frame. To achieve that goal, supply chain management should closely coordinate so can decrease number of channel inventory, eliminate bottlenecks, compress periods and eliminate quality issue.

In context of data analysis in PKT, in accordance with the mandate of the 1945 Constitution, Indonesia must maintain food security, which is a human right. The definition of food security itself is stated in the Law. The Strategy to Maintain National Food Security is regulated in the law. To realize national food security, PT PKT as a subsidiary of PT PI which is a state-owned holding company has the mandate to fulfill the demand for subsidized fertilizers from the government. To find out the extent of PT PKT's ability and to know future strategic policies.

d. Predective Analysis

Predictive analytics is commonly used in statistical and analytics. It is used to predicts the future by analyzing current and historical data. With this approach, the future events and behavior can be predicted based on design models of predictive analytics. Predictive analysis come up with

score given by mostly predictive analytics models which indicates likelihood of occurrence. Historical and transactional data patterns are explored by to find out the solution for many business and science problems. With the increase in attention towards decision support solutions, predictive analytics models have dominated in this field. In this paper, we will present a review of process, techniques, and applications of predictive analytics (Zhai et al., 2020)

Demand Forecasting is one of predictive analysis that is essential in making production decisions. Demand forecasting accuracy affects supply chain management and can reduce its costs. The development of information technology, especially machine learning, has many benefits in many industrial sectors. The development of technology in forecasting can produce better accuracy. The use of machine learning in demand forecasting is in various industrial sectors ranging from small-scale industry to large-scale industry (Dalimunthe et al., 2023)

METHODOLOGY

A. Data Understanding

The data is retrieved from the Material Management Module in the SAP application using transaction code MB51. Subsequently, data is exported from the SAP application to Excel format and then filtered according to the business process transactions conducted (Purchase, Sales, and Stock Raw Material). To carry out the analysis of Purchasing, Sales, and Product Planning of PKT, we will need the data sets listed below to help with the analysis process.

Purchase (Record all purchase history)

Field	**Data type**	**Description**
Posting_Date	datetime	When the purchasing happened
Purchase_Order	Int	This will be used as the purchase id

Quantity	Int	The qty of raw material purchased
Unit_of_Entry	varchar	Units of qty of raw material (Ex: TON)
Material_Description	varchar	Types of raw material purchased
Storage_Location	Varchar	Id for which warehouse material will be stored

Sales (Record all sales history)

Field	**Data type**	**Description**
Posting_Date	datetime	When the sales happened

Quantity	int	How product much sold
Unit_of_Entry	varchar	Unit of qty product sold (TON)
Material_Description	varchar	Types of raw material purchased
Storage_Location	Varchar	Id for which warehouse material sold

Stock (Record in/out/current stock | will be used to fulfill production planning)

Field	**Data type**	**Description**
Material_Description	varchar	Types of raw material purchased
Quantity	int	How product much sold
Unit_of_Entry	varchar	Unit of qty product sold (TON)

Storage_Location	Varchar	Id for which warehouse material sold
Capacity	int	Max qty warehouse can store

B. Data Preprocessing

From the sets of data above, we will need to process and merge the tables to help visualize the trend line from past data to know and predict the trends for inventory stock level status :

Field	**Data type**	**Description**
Clay Qty (TON)	Integer	Numbers of quantity Clay in TON unit
KCL Qty (TON)	Integer	Numbers of quantity KCL in TON unit
DAP Qty (TON)	Integer	Numbers of quantity DAP in TON unit
Season	Integer	Season occurred when NPK produced 1) Dry Season 2) Growing Season
NPK Sales Qty (TON)	Integer	Numbers of sales quantity NPK Product

After all required data merged, we need to split our dataset into data training and data testing in Rapidminer. We'll be using a specific operator (Generalized linear models) for this purpose. GLM takes Dataset as its input and delivers the subsets of that ExampleSet through its output ports. The number of subsets (or partitions) and the relative size of each partition are specified through the partition's parameter. The sampling type parameter decides how the examples should be shuffled in the resultant partitions.

In our case, there are a total of 149 lines in out dataset. We're trying to aim for 70-30 split on our data for training and testing purposes. That means 70% of our data which equal to 104 lines will be used for data training purposes and the remaining 30% of our data which equal to 45 lines will be reserved for testing.

C. Data Modelling

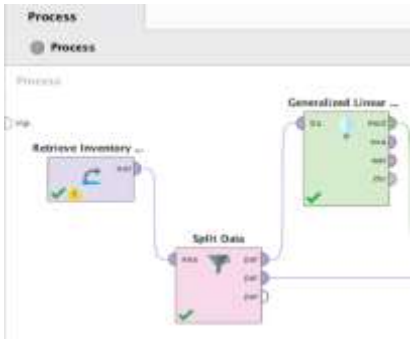
Since inventory level stock identified as Label, this model can be considered as supervised learning. From the preprocessed data above, we can analyze:

- Trends for when raw material level is danger, minimum, average and maximum to fulfill sales demand of NPK product.
- To find sales pattern and demand to arrange purchasing strategy to ensure fulfilment for those demands.
- To find specific scenario when raw material is in Danger level (very low).

For data modelling, we use data mining tools called Rapidminer. RapidMiner is a popular and pretty reliable open-source platform especially for machine learning. Its user-friendly interface make it an ideal choice for a diverse range of users, including data scientists, analysts, and even those without deep programming expertise. This comprehensive platform offers an extensive suite of tools and functionalities, specifically tailored for tasks such as data analytics, machine learning, and data mining. It excels in providing a streamlined and efficient pathway for users to access,

prepare, model, and ultimately deploy data-driven solutions, while significantly reducing the reliance on complex coding or programming skills.

First operator used in this model is Generalized linear models (GLMs), an extension of traditional linear models. This algorithm fits generalized linear models to the data by maximizing the log-likelihood. The elastic net penalty can be used for parameter regularization. The model fitting computation is parallel, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients.



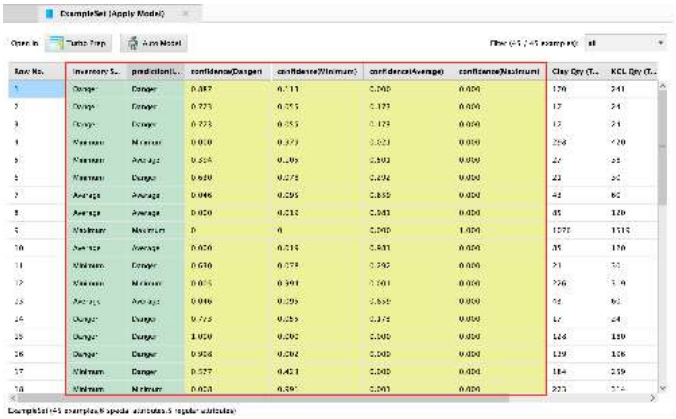
This operator produces the desired number of subsets of the given ExampleSet. The ExampleSet is partitioned into subsets according to the specified relative sizes. Here is the result of GLMs using our data training :

Generalized Linear Model (Generalized Linear Model) | PerformanceVector (Performance)

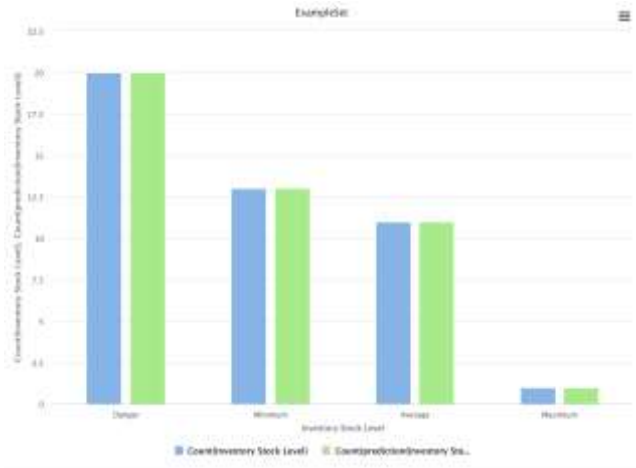
Attribute	Coefficient Da...	Coefficient Ave...	Coefficient Ma...	Coefficient Min...	Std. Coefficient...	Std. Coefficient...	Std. Coefficient...	Std. Coefficient...
Clay Qty (TON)	-0.030	0.014	0.020	0.002	-2.997	1.344	2.021	0.226
KCL Qty (TON)	-0.021	0.009	0.014	0.002	-2.992	1.252	2.010	0.276
DAP Qty (TON)	-0.010	0.004	0.006	0.001	-2.988	1.245	2.019	0.279
Season	19.429	-16.704	0	-0.215	9.087	-7.812	0	-0.100
NPK Sales Qty (T...	-0.026	0.011	0.017	0.002	-2.991	1.259	2.019	0.272
Intercept	-16.444	15.252	-34.400	-1.692	-1.550	-5.272	-22.401	-0.487

In regression analysis, coefficients represent the values associated with independent variables in a regression equation. For example, in linear regression, you have coefficients that represent the relationship between the independent variables and the dependent variable. These coefficients tell you how much the dependent variable is expected to change for a one-unit change in the corresponding independent variable while holding other variables constant.

According to this result, the coefficient of each label value is quite good. Each variable input has relationship to define inventory level stock :



According to apply model result, we can see that prediction of Inventory Level Stock is quite the same with the actual value. There are confidence level for each value of inventory level the prediction will be the highest.



RESULT AND DISCUSSION

To evaluate the model, we use Performance (classification) operator which is used for statistical performance evaluation of classification tasks. This operator delivers a list of performance criteria values of the classification task. There are many criteria parameters to define performance of models. But in this case, we only use 2 common parameter, accuracy and root mean squared error.

According to performance result, prediction accuracy of this model is pretty high, 91,11% means that predictive value is almost 100% correct.

In data mining, a lower RMSE is generally better. According to result in rapidminer, this model has 0.279 root mean squared error indicates that, on average, the model's predictions are relatively close to the actual values. This suggests that the model is making reasonably accurate predictions.

From business perspective, PKT can use this model for several benefits :

1. Scenario Analysis : consider running scenario analyses to based sales forecast data or prepare for potential disruptions or changes in demand, such as the impact of a new competitor, market trends, or unexpected events (e.g., pandemics, natural disasters).
2. Forecasting: Once your model is trained and validated, you can use it to generate sales demand forecasts. These forecasts will provide estimates of future sales based on historical patterns and relevant features.
3. Real-time Monitoring: Continuously monitor the actual sales against the forecasts and adjust production schedules and inventory levels as needed. This real-time feedback loop allows for dynamic adjustments based on changing market conditions.

CONCLUSIONS

The CRISP-DM (Cross Industry Standard Process for Data Mining) provides a structured approach to data mining which helps in making the projects more manageable and repeatable. It also helps in ensuring that the data mining goals are aligned with the business objectives and that the models created are reliable and effective. Overall, the CRISP-DM process model aims to make data mining projects more efficient and successful (Schröer et al., 2021). CRISP-DM can help speed up developing data analysis and data mining projects. It provides a structured and systematic approach to the entire data mining process, from understanding business objectives to deploying and monitoring models. The use of machine learning in demand forecasting is in various industrial

sectors ranging from small-scale industry to large-scale industry (Dalimunthe et al., 2023). According to result in rapidminer, this model has 0.279 root mean squared error indicates that, on average, the model's predictions are close to the actual values. prediction accuracy of Generalized linear models is 91,11%. These forecasts will provide estimates of future sales based on historical patterns and relevant features.

REFERENCES

Dalimunthe, S. B., Sinulingga, S., & Ginting, R. (2023). JSTI Jurnal Sistem Teknik Industri Implementation Of Machine Learning in Demand Forecasting: A Review of Method Used in Demand Forecasting with Machine Learning. *Jurnal Sistem Teknik Industri (JSTI)*, *25*(1), 2023. https://doi.org/10.32734/jsti

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 5, Issue 2). http://jurnal.polibatam.ac.id/index.php/JAIC

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). 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

Parate, A. (n.d.). *Integrating Crisp DM Methodology for a Business Using Tableau Visualization Critical Assessment of Circular Economy Regarding Waste Reduction and Optimal Use of Resources View project Marketing Strategies in the textile industry to influence Gen Y consumers View project*. https://doi.org/10.13140/RG.2.2.36619.31520

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Wirth, R., & Hipp, J. (n.d.). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. In *Computers and Electronics in Agriculture* (Vol. 170). Elsevier B.V. https://doi.org/10.1016/j.compag.2020.105256