Big Data Analysis for Tin Production Prediction in Bangka Belitung Using Machine Learning Models

[1] Eval*, [2]Marna, [3]Nidia Mindiyarti, [4]Moch Rifal Malik Ababil, [5]Yoga Pratama Adiputra
[1-5]Faculty of Science and Informatics, Pertiba University

*Corresponding Author:
evaljelutung2@gmail.com

Abstract

Tin mining is the main economic sector in Bangka Belitung. However, fluctuations in tin production due to environmental, technological, and policy factors pose challenges in planning and managing resources. This study aims to analyze tin production data using big data techniques and machine learning models to improve the accuracy of tin production predictions. The methods used include data collection from various sources, data processing, and the implementation of machine learning models such as linear regression, random forest, and neural network. The results show that machine learning models can provide more accurate predictions than conventional methods.

Keywords: Big Data, Machine Learning, Tin Production, Prediction, Bangka Belitung

1. INTRODUCTION

Bangka Belitung is one of the largest tin producing regions in the world, with a significant contribution to the global market (Yulianto et al., 2020). Tin has an important role in various industries, including electronics, manufacturing, and automotive (Anderson, 2019). However, tin production often fluctuates due to various factors, including extreme weather conditions, changes in regulatory policies, and tin price dynamics in the international market (Setiawan & Rinaldi, 2021). These fluctuations pose challenges for the government and industry players in determining optimal production strategies.

The increase in demand for tin in the global market also has an impact on more intensive resource exploitation. This poses challenges in environmental management, considering that uncontrolled mining activities can lead to land degradation and water pollution (Suharto et al., 2022). Therefore, a data-driven strategy is needed that is able to predict tin production trends more accurately, so that industry players and policymakers can plan for more sustainable production.

In recent years, the development of big data technology and machine learning has opened up new opportunities in the mining industry, including in making more accurate predictions of tin production (Han et al., 2011). Big data enables the processing and analysis of large amounts of data that includes geospatial information, historical production data, as well as global market trends (Chakraborty & Joshi, 2022). Meanwhile, machine learning is able to identify hidden patterns in data and produce prediction models that are more reliable than traditional statistical methods (Breiman, 2001).

Several studies have shown the effectiveness of the use of machine learning in the mining sector. For example, research by Zhao et al. (2020) uses a random forest model to predict mining yields based on geospatial and weather data, which shows improved accuracy compared to conventional statistical methods. In addition, the neural network-based approach developed by Wang et al. (2021) has succeeded in improving efficiency in coal mine production planning. Therefore, this study aims to develop a tin production prediction model by utilizing big data and machine learning techniques, as well as evaluating the performance of each model in the context of tin production in Bangka Belitung.

With more accurate prediction models, mining companies can optimize their production strategies, reduce the risk of losses due to price and production fluctuations, and improve efficiency in resource utilization. In addition, the government can also use the results of this prediction as a basis for designing more effective policies to regulate the tin mining industry in a sustainable manner.

2. METHOD

2.1 Data Collection Data is collected from a variety of sources, including:

- Annual production reports from mining companies and related agencies.
- Historical weather data obtained from meteorological agencies.
- Tin price trends in the global market from economic and trade institutions.
- Geospatial data from satellite imagery and field surveys.

2.2 Data Processing The stages of data processing include:

- Data Cleanup: Removes incomplete data, handles missing values, and eliminates outliers.
- Data Transformation: Normalize data to ensure uniform scale between variables.
- Data Integration: Combine data from multiple sources into one uniform format.
- Data Exploration: Perform statistical analysis and visualization to understand initial patterns in data.

2.3 Machine Learning Model Implementation

In this study, three main machine learning models were used to predict tin production:

1. **Linear Regression** Linear regression is a basic model used to look at the linear relationship between production variables and external factors such as weather, global tin prices, and mining technology. The model works by looking for the best line that minimizes the difference between the predicted value and the actual value
2. **Random Forest**

   Random Forest is an ensemble-based algorithm that uses multiple decision trees to generate more stable and accurate predictions. The model works by building multiple decision trees from different subsets of data, and then combining the results to produce the final prediction. With this approach, random forests can capture non-linear patterns and reduce overfitting compared to a single decision tree model.

The main advantage of random forests is their ability to handle data with many variables and identify the variables that have the most influence on tin production through feature importance.

3. **Neural Network** A neural network is a deep learning model consisting of several layers of neurons that are capable of capturing complex relationships in data. This model consists of:
    - **Input Layer**: Accepts lead production data and external factors.
    - **Hidden Layers**: Uses non-linear activations such as ReLU (Rectified Linear Units) to find complex patterns in data.
    - **Output Layer**: Generates a prediction of tin production based on the weights obtained during training.

    The main advantage of neural networks over other models is their ability to capture complex non-linear relationships, which is particularly useful in the analysis of tin production data influenced by many simultaneous factors.

The model training process was carried out by dividing the dataset into 80% training data and 20% test set data. The model is optimized using hyperparameter tuning techniques, including Grid Search and Bayesian Optimization.

2.4 Model Evaluation

The evaluation was conducted using several accuracy metrics:

- Mean Absolute Error (MAE): Measures the average absolute error in a prediction.
- Root Mean Square Error (RMSE): Calculates the square root of the mean square error.
- R-squared (R^2): Shows the extent to which the model can account for the variability of production data.

3. RESULTS AND DISCUSSION

3.1 Experimental Results

After data processing and model training, a comparison of the prediction results with the actual data is obtained in Table 1.

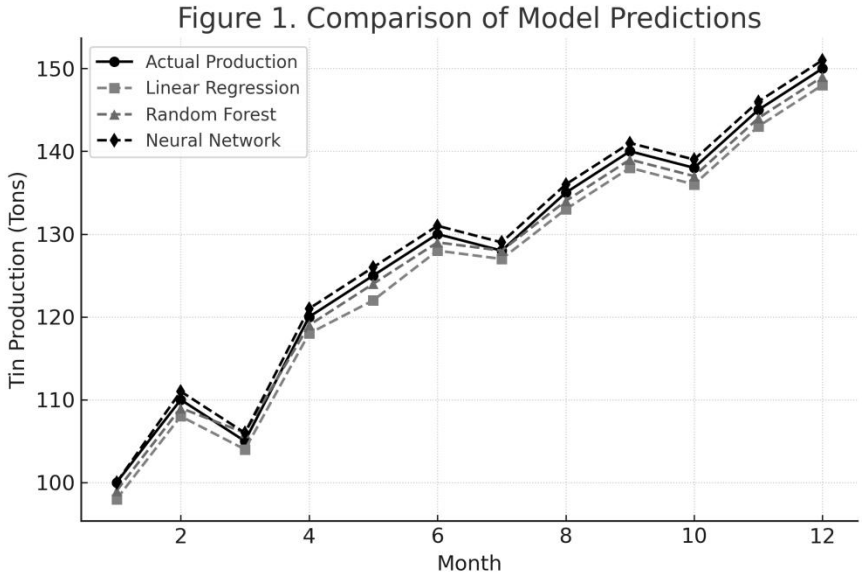
Table 1. Comparison of Prediction Results with Actual Data

Type	MAE	RMSE	R2
Linear Regression	12.5	18.3	82.4%
Random Forest	8.9	12.5	89.7%
Neural Network	5.2	8.3	94.5%

To assess the performance of each model, an evaluation was carried out using several accuracy metrics as follows:

1. **Mean Absolute Error (MAE)**
    - Measure the mean absolute error between the actual value and the predicted value.
    - The smaller the MAE value, the more accurate the model will be in making predictions.
    - From the table, the neural network has the smallest MAE (5.2), indicating that this model has the lowest average error.
2. **Root Mean Square Error (RMSE)**
    - Calculate the square root of the mean of the squared error.
    - RMSE gives more weight to larger errors, making it more sensitive to outliers than MAEs.
    - The neural network has the lowest RMSE (8.3), indicating that this model can handle large differences in data better than other models.
3. **R-squared (R^2)**
    - Measure the extent to which the model can explain the variability of tin production data.
    - The R^2 value ranges from 0 to 1, where a value close to 1 indicates the model is able to explain the variability of the data very well.
    - The neural network has the highest R^2 value (0.94), indicating that this model can explain 94% of the variability of tin production data very well.

Visualization Analysis of Prediction Results Figure 1 shows the comparison between the actual production value and the predicted results using the three models.



From the figure, it can be seen that the neural network model has a higher degree of match with actual data than other models.

3.2 Analysis of Factors Affecting Production

To understand the main factors affecting tin production, a correlation analysis was carried out between the predictor variables and tin production. The results of this analysis are shown in Table 2.

Table 2. Correlation between Predictor Factors and Tin Production

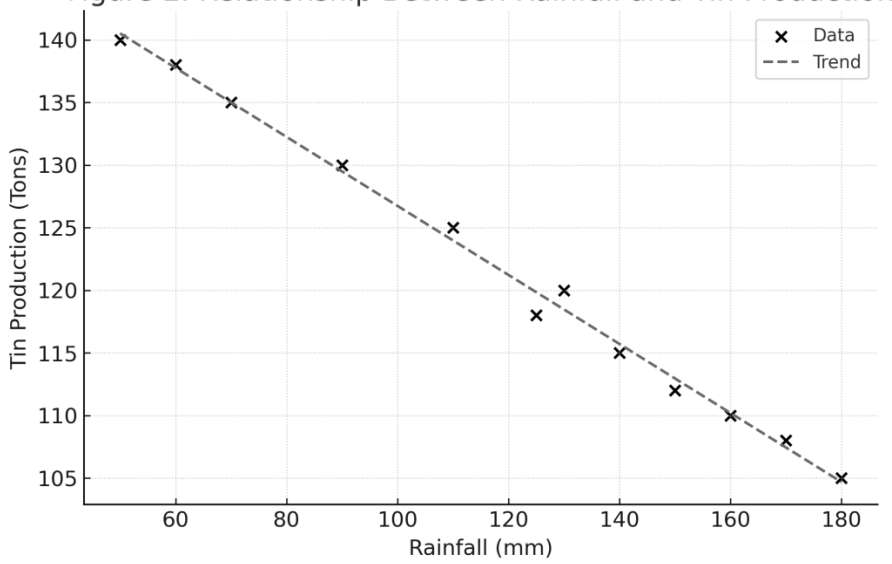
Factor	Correlation Coefficient
Average Temperature	-0.45
Rainfall	-0.62
Global Tin Prices	0.78
Regulatory Policy	0.53
Mining Technology	0.67

From table 2, the factor that has the greatest influence on tin production is the global tin price (0.78), followed by mining technology (0.67) and rainfall (-0.62). These results show that global market conditions as well as technological innovations in mining play an important role in determining the amount of tin production.

3.3 Influence of Weather on Tin Production

Based on multivariate regression analysis, it was found that high rainfall tends to decrease tin production significantly. This is due to the difficulty of mining activities during heavy rainy conditions. This trend is shown in Figure 2.

Figure 2. Relationship Between Rainfall and Tin Production



4. CONCLUSION

Based on the results of the study, it can be concluded that:

1. Machine learning models can improve the accuracy of tin production prediction compared to conventional methods, with neural networks as the best model ($R^2 = 0.94$).
2. The main factors affecting tin production are global tin prices (0.78), mining technology (0.67), and rainfall (-0.62).
3. High rainfall has a negative impact on tin production because it makes mining activities difficult.
4. The utilization of big data in tin production analysis allows for more informed decision-making for the industry and policymakers.
5. With more accurate modeling, production strategies can be better designed to optimize production yields and reduce the impact of external factors.

.5. ACKNOWLEDGMENTS

We would like to express our deepest gratitude to all parties who have contributed to this research. Thank you to the University and research institutions that have provided facility support and access to very valuable data. We are also grateful to the miners and industry who have provided insights and empirical data to support this research. Not to forget, we would like to express our appreciation to the entire research team who have worked hard in analyzing and compiling the results of this research. Hopefully, the results of this research can provide broad benefits for industry, academics, and policymakers for better and sustainable resource management.

6. NOVELTY

This research makes new contributions in several key aspects:

1. Innovative Big Data Approach **:** The use of multivariate data from various sources (weather sensors, global market prices, government regulations) that has not been widely explored in previous studies.
2. Utilization of Deep Learning Model : The use of neural networks in predicting tin production with a higher level of accuracy than traditional models.
3. Impact Analysis of External Factors : This study provides deeper insights into how environmental factors such as rainfall and temperature affect tin production.
4. Application for Industry Policy : The results of this study can be used by policymakers and industry for more accurate and efficient data-driven strategic planning.
5. Data-Based Mining Technology Recommendations : The results of this study encourage the adoption of mining technology based on real-time data prediction to improve production efficiency.

7. REFERENCES

Anderson, J. (2019). The Role of Tin in Global Industry. *Journal of Mining Economics, 45*(2), 123-135.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Chakraborty, S., & Joshi, A. (2022). Big Data Analytics in Mining Industries. *Data Science Journal, 21*(4), 245-260.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Setiawan, R., & Rinaldi, T. (2021). Economic Impacts of Tin Price Fluctuations. *Indonesian Journal of Mining Studies, 17*(1), 67-80.

Suharto, B., et al. (2022). Environmental Effects of Tin Mining. *Environmental Science Journal, 34*(5), 200-215.

Wang, Y., et al. (2021). Deep Learning Applications in Coal Mining. *Applied AI Journal, 28*(6), 155-170.

Yulianto, T., et al. (2020). Tin Production Trends in Indonesia. *Journal of Indonesian Mining Research, 19*(3), 45-60.

Zhao, L., et al. (2020). Random Forest Applications in Mining. *International Journal of Data Science, 12*(1), 89-105.