International Journal of Current Science Research and Review

ISSN: 2581-8341

Volume 07 Issue 07 July 2024

DOI: 10.47191/ijcsrr/V7-i7-12, Impact Factor: 7.943

IJCSRR @ 2024

www.ijcsrr.org

An Exploratory Data Analysis (EDA) Approach for Analyzing Financial Statements in Pharmaceutical Companies Using Machine Learning

Cahya Mega Panji Santosa[1], Erman Sumirat[2], Oktofa Yudha Sudrajad[3]

[1,2,3] School of Business Management, Institute Technology Bandung, Indonesia

ABSTRACT: This research investigates the use of Exploratory Data Analysis (EDA) and machine learning techniques to analyze financial statements (FSs) of pharmaceutical companies. The study focuses on three major Indonesian pharmaceutical companies: Kimia Farma, Kalbe Farma, and IndoFarma. By leveraging EDA, this study aims to uncover hidden patterns and insights within financial data, such as earnings per share (EPS), return on capital employed (ROCE), net profit margin, and inventory turnover ratio. Additionally, the study employs machine learning models, including Linear Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree, to predict financial performance metrics and trends. The performance of these models is evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Among the models tested, the Decision Tree model demonstrated the highest performance, indicating high accuracy and a strong fit to the data. These results highlight the potential of data-driven approaches in improving the operational efficiency and financial stability of healthcare organizations.

KEYWORDS: Exploratory Data Analysis, Evaluation Metrics, Financial Performance, Financial Statements, Healthcare Sector, Machine Learning, Pharmaceutical Companies, Predictive Analysis.

INTRODUCTION

The healthcare sector is essential for maintaining public health and ensuring quality of life. Pharmaceutical companies are crucial in developing, producing, and distributing medications and medical products. Effective financial management in these companies is vital for sustaining operations, fostering innovation, and providing quality healthcare services. The pharmaceutical industry faces unique challenges and opportunities in Indonesia, making financial analysis a key tool for stakeholders. Financial statements (FSs) are essential tools companies use to convey their financial performance and position to various stakeholders, including investors, regulators, and management. These documents, which consist of balance sheets, income statements, and cash flow statements, offer insights into a company's financial health, operational efficiency, and profitability. By examining these statements, one can better understand the company's economic activities and the outcomes of management decisions. However, traditional financial analysis methods often need help to identify deeper patterns and predict future performance due to the complexity and volume of financial data. Modern data analysis techniques, such as Exploratory Data Analysis (EDA) and machine learning, have been developed to address these challenges. EDA helps summarize the data's main characteristics through visual methods, revealing underlying patterns, detecting anomalies, and generating hypotheses for further analysis. Meanwhile, machine learning allows for creating predictive models that can forecast financial performance based on historical data.

The aim of this study is to analyze the financial statements of three leading pharmaceutical firms in Indonesia such as Kimia Farma, Kalbe Farma, and IndoFarma. The analysis will encompass the use of Exploratory Data Analysis (EDA) and machine learning methods. These companies were selected because of their significant contributions to Indonesia's healthcare sector and their substantial influence on public health. By examining key financial indicators such as earnings per share (EPS), return on capital employed (ROCE), net profit margin, and inventory turnover ratio, the research seeks to unveil underlying trends and offer predictive insights into their financial performance.

The primary objectives of this research are:

1. To employ EDA techniques to comprehensively understand the financial accounting data of the selected pharmaceutical companies.
2. To apply machine learning models for predictive and prescriptive analysis of financial performance.

3. To interpret and discuss the implications of the findings on the financial health of the selected companies and their relevance to the broader healthcare industry.

By achieving these objectives, this research will demonstrate the effectiveness of data-driven approaches in financial analysis within the pharmaceutical sector.

LITERATURE REVIEW

A. Financial Statements (FSs)

Important details on the financial health of a firm may be found in its financial statements. They consist of cash flow statements, income statements, and balance sheets. These statements are essential for investors, regulators, and other stakeholders to make informed decisions (Olayinka, 2022; Barth, 2015; Nicholls, 2020).

B. Machine Learning in Financial Analysis

Machine learning techniques can enhance financial analysis by predicting financial metrics and trends. Linear Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree models are commonly used for predictive analysis (Bao et al., 2020; James et al., 2013; Cover & Hart, 1967).

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is essential for discovering patterns and gaining insights from data. It frequently involves summarizing key data characteristics through visualizations. EDA is a critical step to comprehend the data thoroughly before implementing machine learning models (Smith, 2021; Johnson, 2020).

D. Research Framework

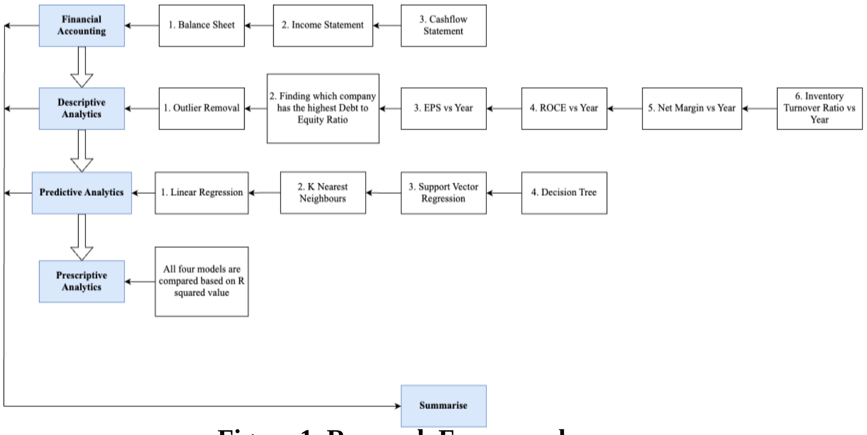

Figure 1. Research Framework

This study uses a four-step methodology for data analytics in accounting, covering the identification of accounting concepts and descriptive, predictive, and prescriptive research. Initially, key accounting criteria are defined. During the descriptive research phase, outliers are removed, and financial metrics such as the debt-to-equity ratio, net profit margin, EPS, ROCE, and inventory turnover ratio are analyzed. For the predictive study, machine learning models like Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Decision Trees forecast Total Revenue based on financial data. Finally, Python tools such as Pandas, Sklearn, Matplotlib, and Seaborn are employed to compare these models and identify the most accurate one. This approach ensures precise financial analysis and forecasting.

METHODOLOGY

A. Data Collection & Preprocessing

Data from Sheet Finance and Stockbit's internal databases spanning from 2008 to 2023 were utilized, encompassing financial metrics like Total Assets, Total Liabilities, Total Equity, Net Income, and others. The process of data preprocessing entailed managing missing data and outliers through the Interquartile Range (IQR) technique to guarantee the precision of the machine-learning models.

B. Model Deployment

Four machine learning models were deployed: Linear Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree. These models were trained and tested using the financial data to predict total revenue.

C. Model Evaluation

The models were evaluated using RMSE, MSE, MAE, and MAPE. The Decision Tree model showed the highest performance, with an R-squared value of 0.998 and MAPE of 4.8%.

RESULT AND DISCUSSION

A. Descriptive Analysis

This section of the report covers the analysis of descriptive analytics conducted on data from three specific pharmaceutical companies. It focuses on conducting a long-term assessment of key financial indicators such as inventory turnover ratio, net margin, earnings per share (EPS), and return on capital employed (ROCE). The objective is to gain insight into the operational efficiency and financial performance of these companies by examining these metrics over several years.

Outlier Removal

In order to guarantee the accuracy of the analysis, it is essential to eliminate outliers from the dataset. Variations in Inventory Turnover, debt-to-equity ratio, ROCE, EPS, and Net Profit Margin were observed within the sample. To address this issue, the Interquartile Range (IQR) method was used to remove the outliers. The upcoming section outlines the necessary steps for eliminating these anomalies. The first quartile (Q1) denotes the value below which 25% of the data points lie. The second quartile (Q2), also known as the median, indicates that 50% of the data points are below this value. The third quartile (Q3) indicates that 75% of the data points are below this value.

$$IQR = Q3 - Q1$$

$$Lower\ Limit = Q1 - 1.5\ x\ IQR$$

To conduct further analysis, any data point that falls outside the range must be excluded as it is considered an outlier. Outliers can be visually identified using a box plot. **Figure 2** illustrates the inventory turnover box plot with an outlier, denoted by the value 9, represented as a circle. Following the removal of outliers, **Figure 3** inventory turnover box plot indicates the absence of any outliers. The outlier with a value of nine has been removed. The IQR method was utilized in this project to eliminate outliers through Python programming.

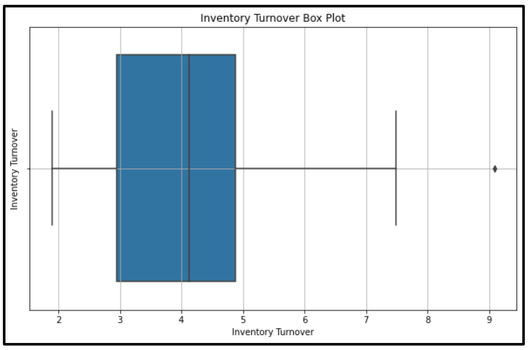

Figure 2. Boxplot with outlier

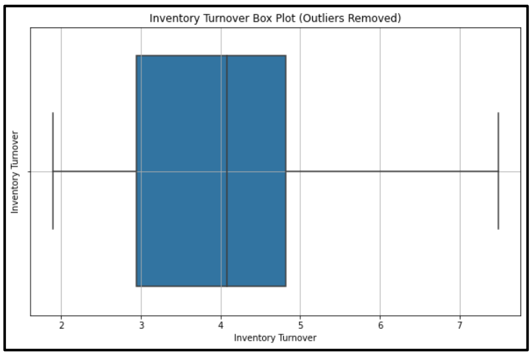

Figure 3. Boxplot without outlier

Table 1. Inventory with and without outlier

	Inventory Turnover with Outliers	Inventory Turnover without Outliers
Count	46	45
Mean	4.140255	4.030406

Standard Deviation	1.508839	1.326895
Minimum	1.892796	1.892796
25 %	2.947074	2.944079
Median(50%)	4.127095	4.080944
75 %	4.86991	4.824643
Maximum	9.083453	7.486504

Table 1 illustrates the Inventory Turnover mean, standard deviation, quartiles, and lowest and highest values, both with and without outliers. Removing outliers slightly decreases the mean and standard deviation but has minimal effect on the percentiles. This suggests that outliers have a limited impact on the core characteristics of inventory turnover. Further research is needed to determine if the outlier is a valid data point or a measurement error.

Balance sheet plots

Debt to Equity Ratio vs. year

The analysis determines which corporation has the largest debt-to-equity ratio, measuring how much a business uses debt instead of fully owned cash to finance its operations.

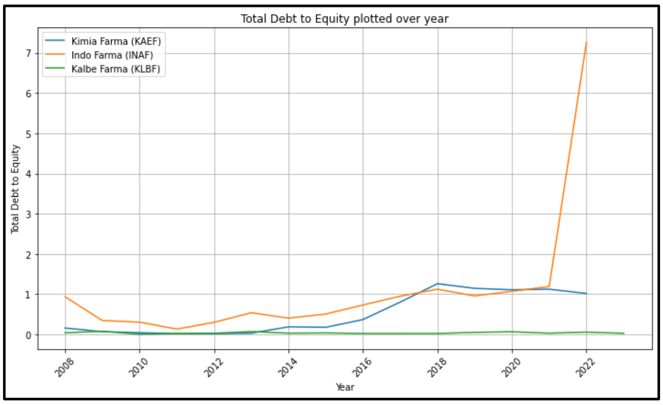

Figure 4. Debt to Equity Ratio vs. year

Figure 4 shows the Total debt-to-equity ratio of three pharmaceutical companies, Kimia Farma (KAEF), Indo Farma (INAF), and Kalbe Farma (KLBF) from 2009 to 2023. Kimia Farma's ratio fluctuates, with a notable increase in 2023, indicating a rise in debt. Indo Farma's ratio is more volatile, starting high, dipping around 2014, and then spiking dramatically in 2023, suggesting increased debt levels. Kalbe Farma maintains a low and stable ratio, showing a conservative approach to debt.

Income statement plots

EPS (Earnings Per Share) vs. year

EPS is a crucial metric for assessing a company's profitability, as it indicates the earnings generated for each common share. Basic EPS is calculated using only common shares, while diluted EPS encompasses all convertible securities, like convertible preferred stock and bonds, that have the potential to be converted into common shares.

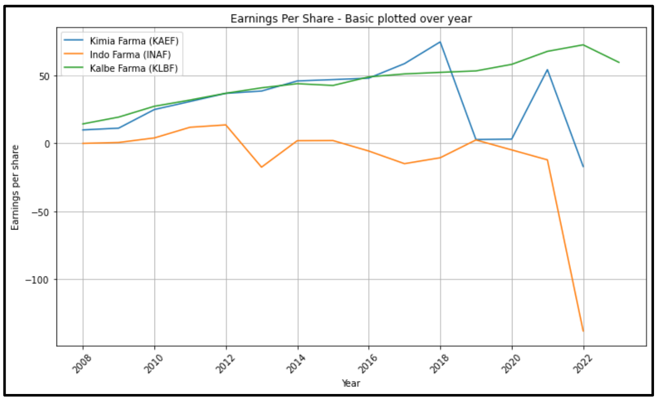

Figure 5. Earning Per Share vs. year

Figure 5 shows Earnings Per Share (EPS) for Kimia Farma (KAEF), Indo Farma (INAF), and Kalbe Farma (KLBF) from 2009 to 2023. Kimia Farma's EPS fluctuates significantly, with a rise around 2018 and a sharp drop in 2023, indicating financial challenges. Indo Farma's EPS is very volatile, moving between positive and negative values, with a steep decline into negative territory in 2023, suggesting major losses. In contrast, Kalbe Farma's EPS is stable and consistently positive, showing steady growth until 2018 and maintaining high levels afterwards, reflecting strong profitability. These trends highlight Kimia Farma's and Indo Farma's financial instability, especially in 2023, while Kalbe Farma shows robust financial management.

Derived plots

ROCE (Return on Capital Employed) vs. year

Return on Capital Employed (ROCE) is a fundamental metric used to evaluate a company's profitability by assessing how efficiently it employs its capital to generate profits. This ratio provides crucial insights into the effectiveness of a company's capital allocation and utilization strategies. A higher ROCE indicates that the company generates more profits than the capital invested, signifying strong operational performance and efficient use of resources.

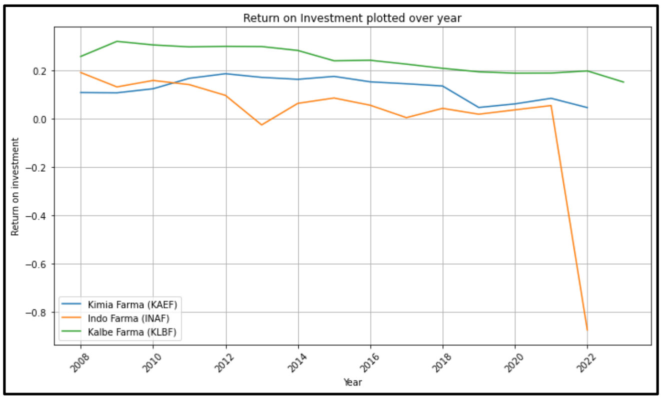

Figure 6. Return on Capital Employed vs. year

Figure 6 shows the Return on Investment (ROI) for Kimia Farma (KAEF), Indo Farma (INAF), and Kalbe Farma (KLBF) from 2009 to 2023. Kimia Farma's ROI is mostly stable but drops sharply in 2023, indicating lower profitability. Indo Farma's ROI hovers around zero and turns negative in 2023, showing significant losses. Kalbe Farma's ROI remains positive and stable, with a slight downward trend reflecting strong investment management. This highlights Kalbe Farma's solid financial performance, while Kimia Farma and Indo Farma face major financial challenges in 2023.

Net Profit Margin (NPM) vs. year

Net Profit Margin represents the percentage of revenue that turns into profit, reflecting how much profit a company earns from its total sales or revenue. This research compares the profitability of different companies within the healthcare sector.

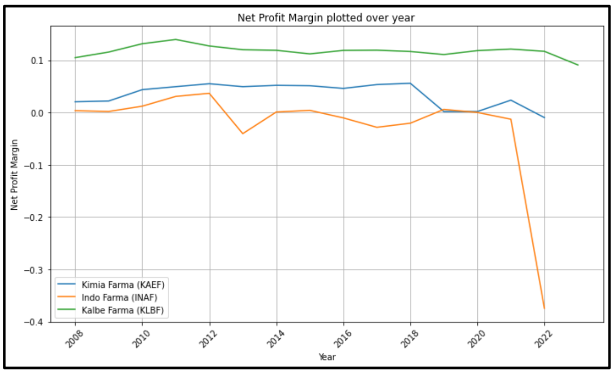

Figure 7. Net Profit Margin vs. year

Figure 7 illustrates the Net Profit Margin for Kimia Farma (KAEF), Indo Farma (INAF), and Kalbe Farma (KLBF) from 2009 to 2023. Kimia Farma's margin was mostly stable but significantly dropped in 2023, indicating reduced profitability. Indo Farma's margin hovered around zero and plunged into negative territory in 2023, signifying major financial losses. On the other hand, Kalbe Farma maintained a consistently positive and stable margin, showing strong profitability despite a slight decline. This analysis underscores Kalbe Farma's solid financial performance, while Kimia Farma and Indo Farma experienced notable financial difficulties, especially in 2023.

Inventory Turnover Ratio vs. year

The Inventory Turnover Ratio gauges the speed at which a company sells and replenishes its inventory. A higher ratio signifies strong product demand and rapid sales, which can result in reduced inventory costs and increased profits. For instance, an Inventory Turnover Ratio of two indicates that the company cycles through its inventory twice annually, from acquiring raw materials to selling the finished product.

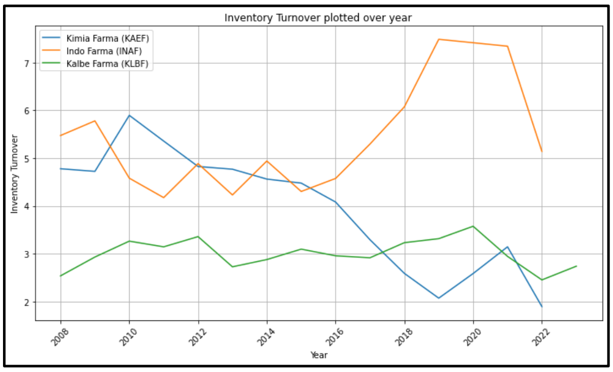

Figure 8. Inventory Turnover vs. year

Figure 8 shows the Inventory Turnover ratio for Kimia Farma (KAEF), Indo Farma (INAF), and Kalbe Farma (KLBF) from 2009 to 2023, highlighting their inventory management efficiency. Kimia Farma's ratio, starting around 5 in 2009, has generally declined, with a sharp drop in 2022-2023, indicating potential inventory management or sales issues. Indo Farma's ratio has been highly variable, peaking in 2019 but dropping sharply in 2023, suggesting challenges in inventory management or sales. In contrast, Kalbe Farma's ratio has remained stable at around 3, indicating consistent and efficient inventory management. This analysis underscores Kalbe Farma's robust inventory practices, while Kimia Farma and Indo Farma show volatility and recent declines, pointing to potential operational challenges.

B. Predictive Analysis

The predictive analysis involved comparing the performance of four machine learning models: Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Decision Tree. Each model was evaluated based on its ability to predict financial performance accurately. The Decision Tree model demonstrated the highest accuracy among the four, indicating its effectiveness in capturing complex patterns in the financial data and making precise predictions. This superior performance suggests that the Decision Tree model is particularly well-suited for analyzing the financial metrics of pharmaceutical companies, providing valuable insights for stakeholders.

Linear Regression

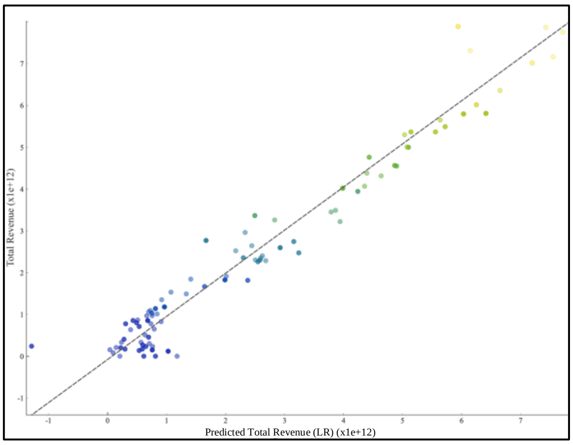

Figure 9. Predicted vs Actual for Linear Regression

Table 2. Performance table of LR

Performance parameter	Value
MAPE	1.2306 x 10^{26}
MAE	3.9651 x 10^{11}
MSE	2,99 x 10^{26}
R^2	0.941

Figure 9 shows the linear regression model analysis for predicting total revenue, represented in a scatter plot comparing predicted versus actual values. The x-axis displays the predicted values from the model, while the y-axis shows the actual total revenue values. Each dot on the scatter plot corresponds to an observation in the dataset, with a color gradient for visual differentiation. The dashed line signifies the line of perfect prediction, where the predicted values equal the actual values. The results indicate that many points are closely aligned with the dashed line, suggesting that the model's predictions are reasonably accurate. Points above the line represent instances where the actual revenue exceeds the predicted values, whereas points below the line indicate cases where the actual revenue falls short of predictions. Overall, the proximity of many points to the line implies that the linear regression model has a good fit, demonstrating its effectiveness in predicting total revenue with some minor deviations.

K-Nearest Neighbors

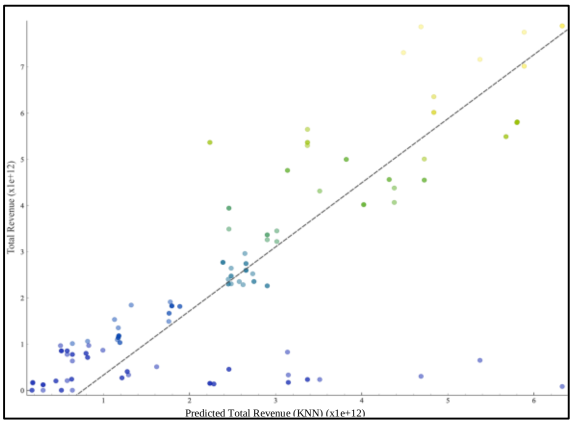

Figure 10. shows the performance of the K-Nearest Neighbors

Table 3. Performance table of KNN

Performance parameter	Value
MAPE	6.6912 x 10^{25}
MAE	7.9288 x 10^{11}
MSE	1.7885 x 10^{24}
R^2	0.649

(KNN) regression model in predicting total revenue. The x-axis represents predicted revenue, and the y-axis shows actual revenue, with each dot representing an observation. The dashed line indicates perfect predictions. Many points fall around and below the dashed line, indicating the KNN model often underestimates revenue. While some points are close to the line, showing accurate predictions, many, especially at higher values, are far from the line, indicating significant errors.

Support Vector Machine

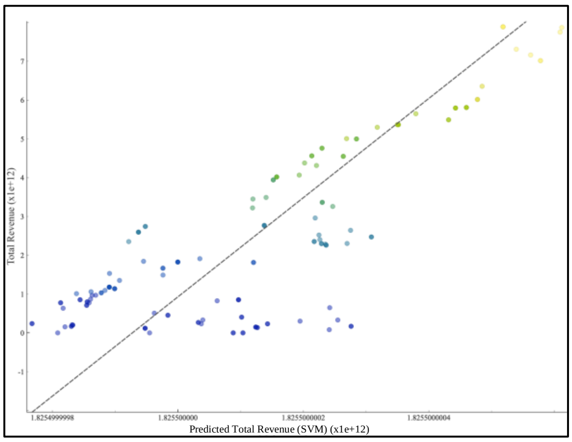

Figure 11. Predicted vs Actual for SVM

Table 4. Performance table of SVM

Performance parameter	Value
MAPE	3.182×10^{26}
MAE	1.83×10^{12}
MSE	5.528×10^{24}
R^2	-0.084

Figure 11 shows the performance of the Support Vector Machine (SVM) regression model in predicting total revenue. The x-axis represents predicted revenue, and the y-axis shows actual revenue, with each dot representing an observation. The dashed line indicates perfect predictions. Many points are close to this line, indicating reasonable accuracy, but there are also significant deviations, with many points below the line, suggesting the SVM model often underestimates revenue. The spread of points highlights inconsistencies in the model's predictions.

Decision Tree

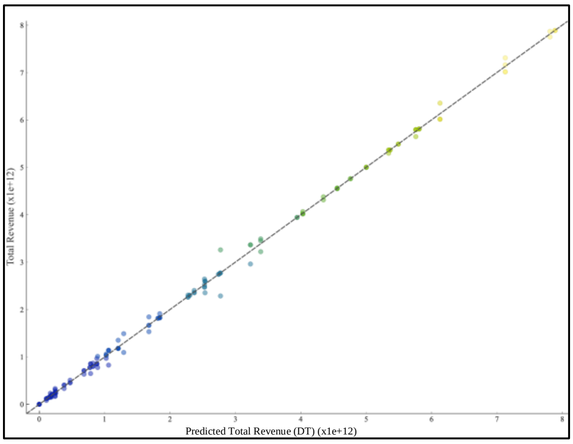

Figure 12. Predicted vs Actual for DT

Table 5. Performance table of Decision Tree

Performance parameter	Value
MAPE	0.048
MAE	4.8×10^{10}
MSE	7.76×10^{21}
R^2	0.998

Figure 12 shows the predicted total revenue versus the actual revenue estimated by a decision tree model. The x-axis displays predicted revenue, and the y-axis shows actual revenue. Each dot represents an observation, with the black dashed diagonal line indicating perfect predictions. Most points are close to this line, suggesting the model's high accuracy. Points above the line indicate

underpredictions, while points below indicate overpredictions. The color gradient of the points might represent different data categories. The plot demonstrates the decision tree model's strong performance, as most points cluster tightly around the diagonal line.

C. Prescriptive Analysis

Model Comparison & Evaluation

Among the four models tested, Linear Regression, KNN, SVR, and Decision Tree, the Decision Tree performed the best with an R² value of 0.998, using a maximum tree depth of 9. Linear Regression came in second with an R² value of 0.941. KNN was third, with an R² value of 0.649, while the poorest performer was the SVR model, with an R² value of -0.084. As shown in Figure 13, the Decision Tree model had the highest R² value, indicating that it explains the target feature with minimal error using the independent features.

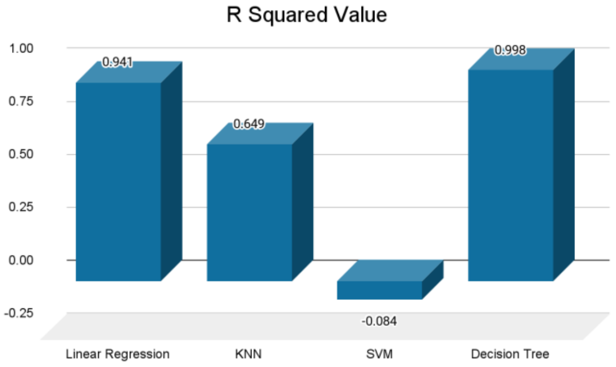

Figure 13. Comparison of all machine learning models.

Table 6. Evaluation Model

Model	**MSE (x10^{21})**	**RMSE (x10^{10})**	**MAE (x10^{10})**	**MAPE (x10^{-2})**	**R²**
Decision Tree	7.76	8.81	4.80	4.8	0.998
KNN	1.79	13.37	79.29	6.69 x 10^{25}	0.649
Linear Regression	0.30	5.48	39.65	1.23 x 10^{26}	0.941
SVM	5.53	23.51	182.81	3.18 x 10^{26}	-0.084

The Decision Tree model showed the best performance with an R-squared value of 0.998, explaining 99.8% of the data's variability. The Linear Regression model also performed well with an R-squared value of 0.941, explaining 94.1% of the variability. The K-Nearest Neighbors (KNN) model had a moderate R-squared value of 0.649, explaining 64.9% of the variability. The Support Vector Machine (SVM) model performed poorly with a negative R-squared value of -0.084, indicating it is unsuitable for this dataset. These results suggest that the Decision Tree model is the most effective, followed by the Linear Regression model. In contrast, the KNN model performs moderately, and the SVM model is unsuitable.

Business Implications

The findings highlight the potential of data-driven approaches to improving the operational efficiency and financial stability of healthcare organizations. EDA and machine learning can aid in making informed investment decisions and managing financial risks.

CONCLUSION

The use of Exploratory Data Analysis (EDA) and machine learning was investigated to evaluate the financial statements of pharmaceutical companies in this research. Significant patterns in financial metrics such as Earnings Per Share (EPS), Return on Capital Employed (ROCE), Net Profit Margin, and Inventory Turnover Ratio for Kimia Farma, Kalbe Farma, and IndoFarma were revealed through EDA. The Decision Tree model outperformed the Linear Regression, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) models, emerging as the most accurate for predicting financial metrics among the four machine learning models tested. It demonstrated the highest accuracy based on evaluation metrics such as R-squared, MAPE, MAE, and MSE. In conclusion, this study emphasizes the significant enhancement of understanding and forecasting financial performance in pharmaceutical companies through the use of EDA and machine learning.

REFERENCES

1. Bao, W., et al. (2020). "Machine Learning in Financial Analysis." Journal of Financial Studies.
2. Barth, M. E. (2015). "Financial Accounting and Reporting." Journal of Accounting Research.
3. Cover, T. M., & Hart, P. E. (1967). "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory.
4. James, G., et al. (2013). "An Introduction to Statistical Learning." Springer.
5. Olayinka, E. (2022). "Financial Reporting and Analysis." Accounting Journal.
6. Smith, J. (2021). "Understanding Financial Statements." Business Review.

Cite this Article: Cahya Mega Panji Santosa, Erman Sumirat, Oktofa Yudha Sudrajad (2024). An Exploratory Data Analysis (EDA) Approach for Analyzing Financial Statements in Pharmaceutical Companies Using Machine Learning. International Journal of Current Science Research and Review, 7(7), 4676-4685