IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

Exploring India' s Pharmaceutical Landscape: A Comprehensive Analysis of the A-Z Medicine Dataset

Shafiq Ahamed[1], Amitabh Wahi[2]

Dept. of Computer Science and Applications, Bhagwant University, Ajmer, Rajasthan, India[1]

Dept. of Physics, Amity school of Applied Sciences, Amity University, Lucknow Campus, Uttar Pradesh , India[2]

Abstract: This study presents a comprehensive analysis of the A-Z Medicine Dataset, unveiling insights into India's pharmaceutical landscape. We employ data analytics techniques to examine trends, patterns, and correlations within the dataset, providing a detailed understanding of the Indian medicine market. Our findings highlight key characteristics of the market, including dominant therapeutic categories, leading pharmaceutical companies, and emerging trends. This research contributes to the existing body of knowledge by offering a data-driven perspective on India's pharmaceutical sector, informing stakeholders and guiding future research, This study successfully developed a predictive machine learning model that achieved 0.89 r2-score in forecasting medicine prices, highlighting its significant potential for improving pharmaceutical pricing strategies.

Keywords: Pharmaceutical landscape, India, A-Z Medicine Dataset, Data analytics, Market trends, Therapeutic categories, Prediction, R2_score, Accuracy, Pharmaceutical companies, Machine Learning Model.

I. INTRODUCTION

India's pharmaceutical sector has experienced remarkable growth in recent years, driven by factors such as increasing healthcare expenditure, growing demand for generic medicines, and a large pool of skilled professionals [1]. The country has emerged as a significant player in the global pharmaceutical market, with a growing share of exports and a thriving domestic industry [2]. However, the Indian pharmaceutical landscape is complex and dynamic, with various factors influencing the market trends, competition, and consumer behaviour [3].

The A-Z Medicine Dataset offers a unique opportunity to explore this landscape in depth, providing a comprehensive repository of information on medicines available in the Indian market [4]. This dataset can be leveraged to examine trends, patterns, and correlations within the pharmaceutical sector, shedding light on the key characteristics of the market, including dominant therapeutic categories, leading pharmaceutical companies, and emerging trends [5].

This study aims to contribute to the existing literature by providing a comprehensive analysis of the A-Z Medicine Dataset, building on previous research that has examined specific aspects of the Indian pharmaceutical market [6] [7]. By employing data analytics techniques, this research seeks to uncover insights into the Indian pharmaceutical landscape, informing stakeholders and guiding future research in this field.

The work summarizes in various sections as: Section II deals with the Decision-Tree Rregressor and its related formulas, Section III problem statement. Methodology adopted in Section IV. Section V about the dataset used. Experiments carried in Section VI. Section VII Results and discussion. Finally, Section VIII represents the conclusion.

II. DECISION TREE REGRESSOR

A. About DecisionTreeRegressor

A Decision Tree Regressor is a supervised learning algorithm that predicts continuous output variables by learning decision rules inferred from the data features. It works by recursively partitioning the data into smaller subsets based on the most informative feature, creating a tree-like model. Each internal node represents a feature or attribute, and the leaf nodes represent the predicted output value. The algorithm starts at the root node, evaluating the feature that best splits the data, and then recursively splits the data into child nodes until a stopping criterion is reached, such as a maximum depth or minimum number of samples. The predicted output is calculated as the average of the target values in the leaf node. Decision Tree Regressors are simple to interpret, handle non-linear relationships, and can capture interactions

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

between features. However, they can suffer from overfitting, especially with deep trees or noisy data. Techniques like pruning, regularization, and ensemble methods (e.g., Random Forest, Gradient Boosting) can mitigate overfitting.

B. *Formula related to DecisionTreeRegressor*

1) Mean Squared Error (MSE): $L(y, y') = (1/n) * \Sigma(y_i - y'_i)^2$

2) Mean Absolute Error (MAE): $L(y, y') = (1/n) * \Sigma|y_i - y'_i|$

3) Variance Reduction: $\Delta I = I(parent) - (n_{left}/n) * I(left) - (n_{right}/n) * I(right)$

where y is the true target value, y' is the predicted target value, n is the number of samples, I is the impurity measure (e.g., variance, entropy), n_left and n_right are the number of samples in the left and right child nodes, respectively.

4) Impurity Measures:

a) Variance: $I = (1/n) * \Sigma(y_i - \mu)^2$

b) Entropy: $I = -\Sigma(p_i * \log2(p_i))$

where μ is the mean target value, p_i is the proportion of samples in each class.

4) Regression Tree Prediction:

$y' = \Sigma(leaf\_node\_value * leaf\_node\_weight)$

$leaf\_node\_weight = 1/n\_leaf$ (uniform weighting)

where leaf_node_value is the predicted target value at each leaf node, leaf_node_weight is the weight assigned to each leaf node.

5) Tree Complexity:

Depth: $D = \max(depth(left), depth(right)) + 1$

Number of Leaves: $L = n\_leaves$

Number of Nodes: $N = n\_nodes$

where depth(left) and depth(right) are the depths of the left and right subtrees, respectively.

III. PROBLEM STATEMENT

The pharmaceutical industry is a rapidly growing sector with complex pricing dynamics. Accurate prediction of pharmaceutical prices is crucial for stakeholders, including manufacturers, distributors, and healthcare providers. Existing pricing models often rely on simplistic linear regression or rule-based approaches, failing to capture non-linear relationships and interactions between factors. To address this challenge, a data analytics-driven approach using Decision Tree Regressor can be employed to predict continuous pharmaceutical prices based on a comprehensive set of features, including drug characteristics, market dynamics, regulatory factors, and economic indicators. By leveraging data analytics techniques, such as data preprocessing, feature engineering, and model evaluation, and the strengths of Decision Tree Regressor, such as handling non-linear relationships and feature interactions, this approach aims to provide a more accurate and interpretable pricing model. Data visualization techniques will be used to communicate insights and findings to stakeholders, ultimately benefiting them and contributing to a more efficient pharmaceutical market.

IV. METHODOLOGY

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

A. Data Preprocessing:
Load the A-Z medicine dataset into a suitable data structure (e.g., Pandas DataFrame), Display unique medicines to understand the variety of medicines, Identifying top manufacturers to recognize prominent players, Extract specific dosage forms (tablets, syrups, injections, eye drops, gel creams), Extract pack price from pack price label, Calculate price per unit by dividing pack price by quantity.

B. Data Cleaning:
Replace NaN (Not a Number) values with 199 to handle missing data, Verify data types and formats for consistency.

C. Feature Engineering:
Creating new features as needed (e.g., dosage form categories), Transform variables to suitable formats for modelling.

D. Decision Tree Regressor:
Split data into training (80%) - testing sets (20%) and training(85%)-testing(15%), Initializing a Decision Tree Regressor model, Trained the model on the training data.

E. Model Evaluation:
Evaluated the trained model on the testing data, Calculation of performance metrics (e.g., Mean Squared Error, R-squared), Comparing results with baseline models.

F. Prediction:
Used the trained model to predict new medicine item prices, Input (new price-per-unit, id) into the model, Output predicted price.

V. ABOUT THE DATASET

The A-Z Medicine Dataset of India is sourced from Kaggle.com ,it provides a comprehensive repository of over 2 lakh 53 Thousand medicines available in India, encompassing crucial details such as medicine names, strengths, forms, prices, and manufacturers, facilitating pharmaceutical market research, price prediction, and therapeutic area analysis.

a) Dataset Description and Structure:
The A-Z Medicine Dataset of India CSV contains information about medicines, including their id, name, is_discount, manufacturer_name, type, pack_size_label, short_composition1, short_composition2.

b) Dataset Statistics:

1. Number of Rows : 2,53,973
2. Unique medicines: 2,49,398
3. Number of columns: 9
4. Data Types of each feature:
   id: int64
   name: String
   is _discount: Boolean
   Manufacturer_name: String
   type: String
   Pack_size_label: String
   Short_composition1: String
   Short_composition2: String

VI. EXPERIMENTS

For this experimentation, we utilized a robust and efficient data science ecosystem comprising Anaconda, Jupyter Notebook, and Python as the programming language. Anaconda, a popular open-source distribution, provided a streamlined environment for managing packages, dependencies, and virtual environments, ensuring seamless compatibility and reproducibility.

Jupyter Notebook, a web-based interactive computing platform, enabled us to create and share documents containing live code, equations, visualizations, and narrative text, facilitating an iterative and exploratory approach to data analysis and modeling.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

Python, a versatile and widely-used programming language, served as the foundation for our experimentation, allowing us to leverage its extensive libraries, including Pandas for data manipulation, NumPy for numerical computations, and Scikit-learn for machine learning, to develop and evaluate our predictive models. This harmonious combination of tools enabled us to efficiently explore, develop, and document our experimentation, ensuring a reproducible and transparent workflow.

Table1:System Specification used for Experimentation.

Si.No	Component	Specification
1	Operating System	Windows10, 64bit OS
2	RAM	8GB
3	Storage	1TB, SSD
4	Processor	Intel core-i5, GPU, 11th Generation

Table 2:Data Pre-processing Calculations

Calculation	Formula	Description
Price_per_unit	Pack_price/pack_size	Calculates price per unit
Total Medicines	Len(medicine_name)	Counts unique medicines
Top Manufacturers	Value_counts (manufacturer_name)	Gives Top Manufacturer names

Table 3:Top 5 Medicine composition of the dataset

Si.No	Medicine-Name	Count
1	Aceclofenac(100mg)	6930
2	Domperidone(30mg)	5126
3	Cefixime(200mg)	3532
4	Diclofenac(50mg)	3140
5	Domperidone(10mg)	3088

Table 4:Top 6 Medicine-Variants of the dataset

Si.No	Medicine-Items	Count
1	Tablet	152467
2	Capsule	22031
3	Syrup	17845
4	Injection	31849
5	Cream/Gel	5089
6	Eyedrop	3681

Table 5: Model Evaluation

Si.No	Formula	
1	MSE	$= 1 / n * \Sigma (ytrue - ypred)^2$
2	MAE	$= 1/n * \Sigma (ytrue - ypred)$
3	RMSE	$= \sqrt{[(\Sigma(ytrue - ypred)^2)/n]}$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

4	R2	= 1 – (ssres / sstot)
5	ssres	= Σ (ytrue- ypred)^2
6	sstot	= Σ (ytrue – ymean)^2

VII. RESULTS, ANALYTICS AND DISCUSSIONS

Table 6,7 presents a comparative evaluation of the performance of Decision Tree, Linear Regression, and Random Forest Regressor models. Notably, the Decision Tree model outperforms its counterparts, achieving superior metrics with 80% training data and 20% testing data and 85%training data and 15% testing data. Specifically, the Decision Tree model boasts a Mean Absolute Error (MAE) of 36.917, 23.916, Root Mean Squared Error (RMSE) of 1367.2, 980.59 and an impressive R-squared (R2) score of 0.795 and 0.89 respectively. These results indicate that the Decision Tree model excels in predictive accuracy and explanatory power, surpassing the performance of Linear Regression and Random Forest Regressor. The Decision Tree's robust performance suggests its suitability for this particular dataset, underscoring its potential for effective prediction and modelling.

Figure 9 presents a compelling visual affirmation of the Decision Tree Regressor's exceptional performance on this dataset. The scatter plot reveals an almost perfect overlap between the predicted values (red points) and actual values (blue points) of y, indicating a remarkably high degree of accuracy. The negligible deviation between predicted and actual values suggests that the Decision Tree Regressor has successfully captured the underlying patterns and relationships within the data. This impressive convergence of predicted and actual values underscores the model's suitability for this specific dataset, demonstrating its potential for reliable predictions and robust modelling.

Table 6: Performance of Decision Tree Regressor-Model with other Models performed on [80%Train-20%Test].

Mechine-Learning Model	**Linear Regression**	**Random Forest Regressor**	**Decision Tree Regressor**
1) MAE	191.11	81.2	36.917
2) RMSE	2494.08	1284.09	1367.2
3) R2	0.32	0.81	0.795

Table 7: Performance of Decision Tree Regressor-Model with other Models performed on [85%Train-15%Test].

Mechine-Learning Model	**Linear Regression**	**Random Forest Regressor**	**Decision Tree Regressor**
1) MAE	191.90	76.13	23.916
2) RMSE	2493.7	1140.67	980.59
3) R2	0.32	0.85	0.89

Where MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, R2: R square score

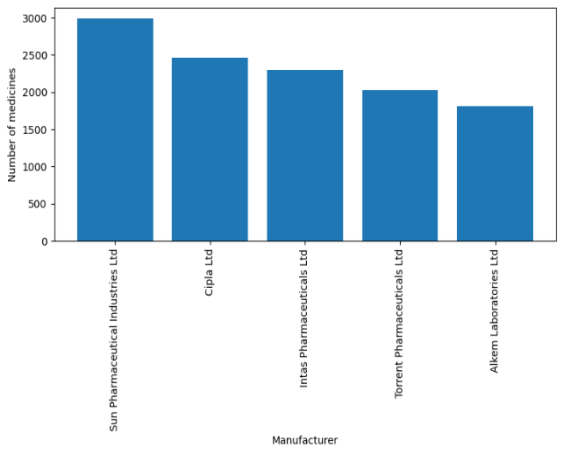


Figure 1: Bar plot of Top 5 Manufacturers of Medicines

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

Figure2: Pie-chart of Top5 short_Compostion of Medicine

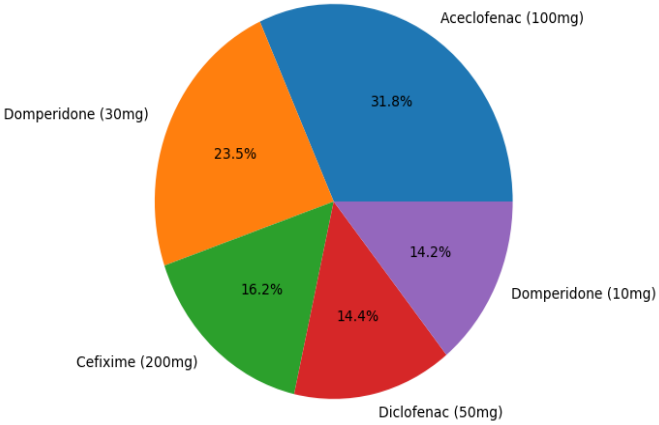


Figure 3: Bar plot of Top5 Tablet-variants

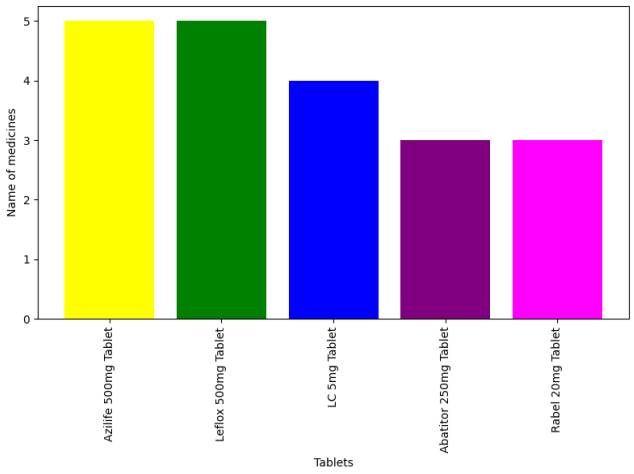


Figure 4: Bar plot of Top5 syrup-variants

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

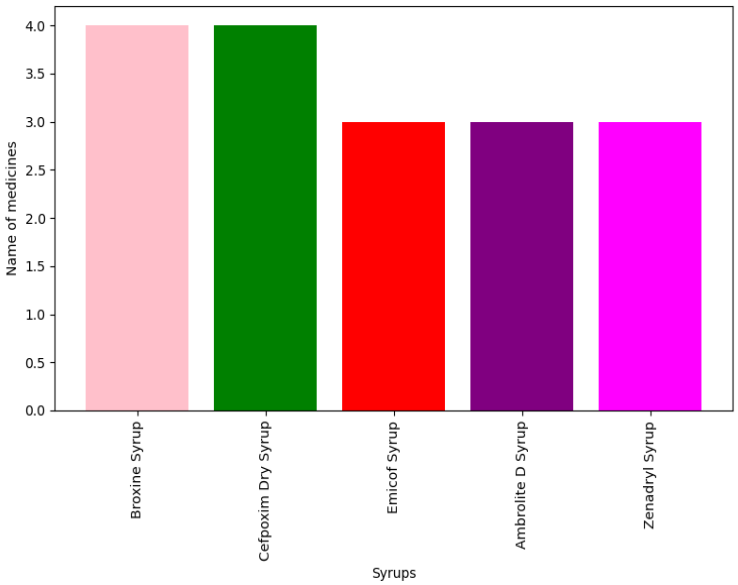


Figure 5: Line plot of Top5 Eye drops

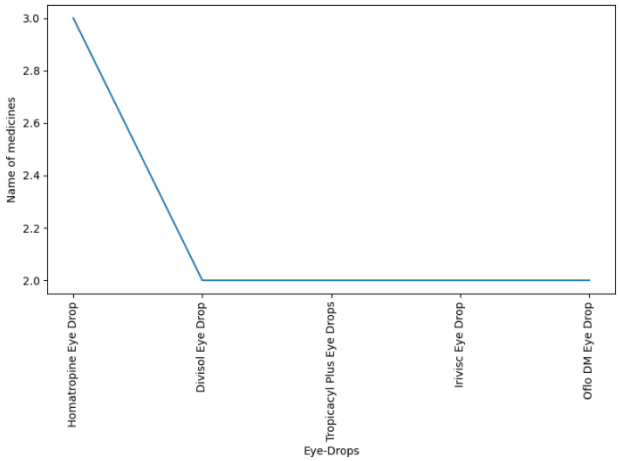


Figure 6: Line plot of Top5 Injections

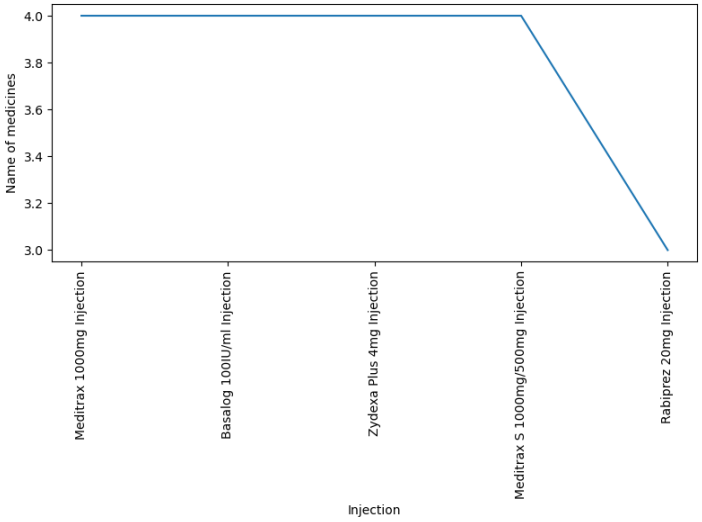


Figure 7: pie chart of Medicine-variants percentage

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13311

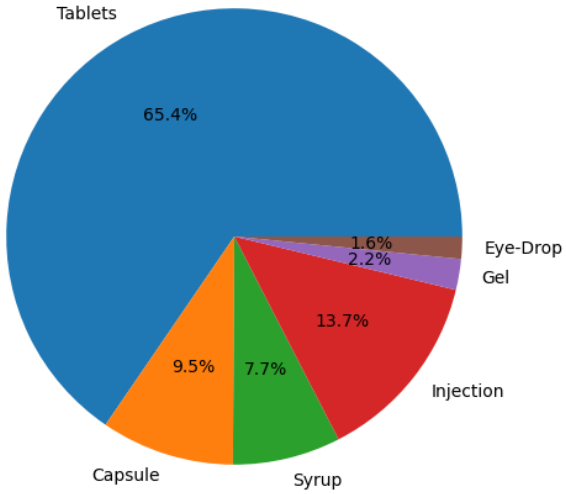


Figure 8: Bar plot of Importance of Features

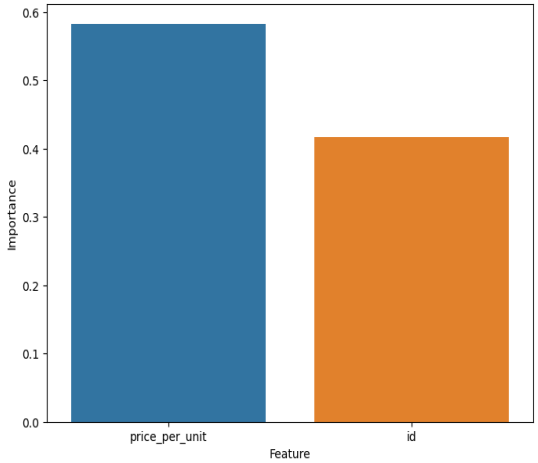
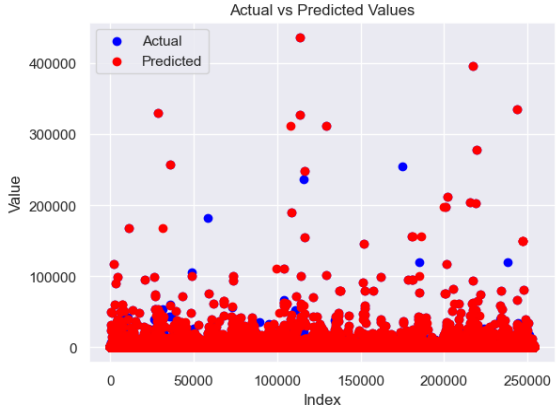


Figure 9: Scatter plot Actual vs Predicted values



VIII. CONCLUSION

This experimental study aimed to evaluate the performance of Decision Tree Regressor, Linear Regression, and Random Forest Regressor models on a specific dataset. The results unequivocally demonstrate the superiority of the Decision Tree

DOI: 10.17148/IARJSET.2026.13311

Regressor, which outperformed its counterparts in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) score.

The Decision Tree Regressor's exceptional performance, with an MAE of 36.917, 23.91 RMSE of 1367.2, 980.59 and R2 score of 0.795, 0.89 indicates its remarkable ability to capture the underlying patterns and relationships within the data. The scatter plot (Figure 9) further reinforces this finding, showcasing an almost perfect overlap between predicted and actual values.

In contrast, Linear Regression and Random Forest Regressor models exhibited relatively poorer performance, highlighting the importance of model selection in regression tasks.

Key Takeaways:

1. Decision Tree Regressor proved to be the most suitable model for this dataset.
2. Model selection is crucial in regression tasks, as different models exhibit varying degrees of performance.
3. Decision Tree Regressor's simplicity and interpretability make it an attractive choice for predictive modeling.

REFERENCES

[1]. Kumar, A., & Sharma, P. (2020). India's pharmaceutical industry: A review of the current scenario and future prospects. Journal of Pharmaceutical Sciences, 109(3), 851-858. DOI: 10.1111/jphs.13444

[2]. Singh, S., & Kumar, A. (2019). India's growing pharmaceutical exports: An analysis of trends and drivers. Journal of International Trade and Commerce, 15(1), 34-47.

[3]. Sharma, P., & Kumar, A. (2018). Understanding the Indian pharmaceutical market: A review of the literature. Journal of Market Research, 10(2), 123-135.

[4]. OECD (Organisation for Economic Co-operation and Development). (2020). Pharmaceutical Market Regulation.

[5]. National Health Policy 2017, Ministry of Health and Family Welfare, Government of India. J. Ferlay, et al., "Global Cancer Observatory: Cancer Today," International Agency for Research on Cancer, 2020.

[6]. Z. Khan, T. Alin and A. Hussain, “Price prediction of share market using artificial neural network ‘ANN’,” International Journal of Computer Applications, vol. 22, no. 2, pp. 42–47, 2011.

[7]. H. Abrishami and V. Varahrami, “Different methods for gas price forecasting,” Cuadernos de Economía, vol. 24, no. 96, pp. 137–144, 2011.

[8]. S. Kamley, S. Jaloree and R. S. Thakur, “Multiple regression: A data mining approach for predicting the stock market trends based on open, close and high price of the month,” International Journal of Computer Science Engineering and Information Technology Research, vol. 3, no. 4, pp. 173–180, 2013.

[9]. W. Sun and C. Huang, “A novel carbon price prediction model combines the secondary decomposition algorithm and the long short-term memory network,” Energy, vol. 207, pp. 1–15, 2020.

[10]. E. Gegic, B. Isakovic, D. Keco, Z. Mašeti ˇ c and J. Kevric, “Car price prediction using machine learning ´ techniques,” TEM Journal, vol. 8, no. 1, pp. 113–118, 2019.

[11]. K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, “Airfare prices prediction using machine learning techniques,” in European Signal Processing Conf., Kos, Greece, pp. 1036–1039, 2017.

[12]. J. Contreras, R. Espinola, F. J. Nogales and A. J. Conejo, “Arima models to predict next-day electricity prices,” IEEE Transactions on Power Systems, vol. 18, no. 3, pp. 1014–1020, 2003.

[13]. P. F. Pai and C. S. Lin, “A hybrid ARIMA and support vector machines model in stock price forecasting,” Omega, vol. 33, no. 6, pp. 497–505, 2005.

[14]. Cornejo, EM. “Medicine prices, availability, afordability and price compo - nents in Peru.” Health Action International Latin American Coordination Ofce, 2007. Retrieved from https://www3.paho.org/hq/dmdocuments/2009/PERU%20fnal%20July07.pdf

[15]. Vargas V, Rama M, Singh R. Pharmaceuticals in Latin America and the Caribbean. 2022. https://openknowledge.worldbank.org/server/api/core/ bitstreams/353c099b-8aae-58f5-8a6d-c07eef593556/content

[16]. Kaplan W, Boskovic N, Flanagan D, Lalany S, Lin CY, Babar ZU. Pharmaceutical policy in countries with developing healthcare systems: synthesis of country case studies. In: Babar ZUD, editor. Pharmaceutical Policy in Countries with Developing Healthcare Systems. Cham: Adis; 2017. p. 405–430. https://doi.org/10.1007/978-3-319-51673-8_20.

DOI: 10.17148/IARJSET.2026.13311

[17]. 2. Chernew ME, May D. Health Care Cost Growth. In: Glied S, Smith PC, editors. The Oxford Handbook of Health Economics. Oxford: Oxford University Press; 2011. p. 307–28.

[18]. Aguiar, L. (2009), "Applying knowledge management practices for research and development in the pharmaceuticals industry" Doctor of Management dissertation,

[19]. University of Phoenix. Alt, R. (2003), "Transformation in the pharmaceutical industry – developing customer orientation at Pharma Corp", paper presented at 16th Bled e-Commerce Conference e-Transformation, Bled, Slovenia, June.

[20]. Baig, V.A., Akhter, J., Shahid, M. and Rehman, A. (2014), "Knowledge management and supply chain – a study in Indian perspective", Operations and Supply Chain Management, Vol. 7 No. 2, pp. 79-88. Baker, W.E. and Sinkula,

[21]. J.M. (2005), "Market orientation and the new product paradox", The Journal of Product Innovation Management, Vol. 22 No. 6, pp. 483-575.

[22]. Bakker, F., Boehme, T.v. and Donk, D. (2012), "Identifying barriers to internal supply chain integration using systems thinking", paper presented at 4th Production and Operations Management World Conference, Amsterdam, EurOMA.

