International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Pharmaceutical Supply Chain Management

Manvi Gour[1], Abhishek Thorat[2], Arnav Singhal[3], Jayesh Ageawal[4] Ajinkya Valanjoo[5], Kusum Kardam[6]

Department of Artificial Intelligence and Data Science
Vivekanand Education Society Institute of Technology, Mumbai, Maharashtra, India

---

Abstract - Supply chain management (SCM) refers to the coordination and oversight of all activities involved in the production, procurement, transformation, and delivery of goods or services from the point of origin to the point of consumption. It encompasses the planning, execution, and control of various processes involved in the flow of materials, information, and finances across the entire supply chain. In this paper, the authors have done a detailed analysis of the pharmaceutical dataset and predicted useful metrics that can be used in the shipping industry. This paper aims to demonstrate the application of machine learning models on the prediction of shipping mode. The model is trained on various features of the pharmaceutical supply chain dataset such as unit price, first-line designation, and delivery dates. Data cleaning and feature engineering were done and multiple models were trained. Out of the models tested, XGboost was found to give the best results for the prediction of weight, shipment modes and delivery date.

***Key Words*:** Supply chain management, Machine Learning models, Shipping Industry, XG Boost

1. INTRODUCTION

Small-scale pharmaceutical companies face challenges in optimizing their supply chain operations. This research investigates developing a predictive model that incorporates various factors to optimize transportation logistics. The model will be trained on historical order data to predict the most likely shipping mode based on factors like order quantity, drug production costs, shipping expenses, and delivery schedules. The model will be evaluated for accuracy, interpretability, and generalizability. Below are the Steps followed:

1. Gather historical order data from small-scale pharmaceutical companies.

2. Clean and prepare the collected data for analysis.

3. Extract relevant features from the data.

4. Train and evaluate various machine learning models to predict shipment mode, delivery date, and weight of the product.

5. Select the best-performing model based on its accuracy, interpretability, and generalizability.

6. Implement the selected model in a real-world setting and evaluate its performance

2. Problem Statement

Small-scale pharmaceutical companies, constituting a pivotal segment of the pharmaceutical industry, confront a notable challenge in their supply chain operations. Specifically, these enterprises grapple with the absence of a comprehensive predictive model capable of accounting for multifarious influential variables and factors such as Sub Classification, Unit of Measure (Per Pack), Scheduled Delivery Date (Year, Month, Day), Delivered to Client Date (Year, Month, Day), Delivery Recorded Date (Year, Month, Day), Line Item Quantity, Line Item Value, Weight (Kilograms), Freight Cost (USD), Pack Price. that bear upon the selection of the most efficient mode of transportation within their intricate supply chains. This conspicuous deficiency in their logistical decision-making toolkit has profound implications for cost optimization, environmental sustainability, and overall competitive positioning within the pharmaceutical sector. Consequently, this research addresses the pressing issue faced by small-scale pharmaceutical firms by delving into the development of a predictive model tailored to their specific operational context, thereby empowering them to make informed and strategic choices regarding transportation modes.

2.3. Existing Solution

Existing methods for predicting optimal transportation modes in small-scale pharmaceutical companies are inadequate. Many of these methods rely on simplistic heuristics, overlooking crucial factors such as sub-classification, unit of measure, scheduled delivery dates, and line item quantity. Consequently, these oversimplified approaches often lead to suboptimal transportation choices, resulting in inefficiencies and cost overruns within supply chains.

Moreover, the absence of a comprehensive, data-driven model exacerbates this problem, leaving companies without the necessary tools to make informed decisions. To address these limitations, there is an urgent need to advance transportation mode prediction methods for small-scale pharmaceutical firms. This requires developing a more sophisticated predictive model that considers the full spectrum of influential factors.

By leveraging such a model, companies can optimize their transportation logistics comprehensively, enhancing cost-effectiveness, sustainability, and overall competitiveness in the dynamic pharmaceutical industry landscape.

This shift towards data-driven decision-making will empower companies to navigate the complexities of supply chain

management more effectively, ultimately driving better outcomes for their businesses.

2.2. RESULTS AND DISCUSSION

The proposed predictive model achieved an accuracy of 96.4% in predicting the most likely shipping mode for small-scale pharmaceutical companies. The model also demonstrated high interpretability, allowing decision-makers to understand the rationale behind its recommendations. Furthermore, the model exhibited good generalizability, performing well on unseen data. These results suggest that the proposed predictive model can effectively optimize transportation logistics in small-scale pharmaceutical companies. By leveraging the model's ability to accurately predict shipping modes, companies can reduce shipping costs, improve supply chain efficiency, and enhance their competitiveness in the pharmaceutical industry. Future research could focus on refining the predictive model by incorporating additional data sources and exploring more advanced machine-learning techniques. Additionally, investigating the impact of external factors such as environmental regulations and technological advancements on transportation mode selection could provide valuable insights for the pharmaceutical industry.

3. LITERATURE SURVEY

Current supply chain management systems address the issue of optimization of the supply chain and the prediction of the delivery methods by existing methods such as historical data analysis, applying basic mathematical models which include optimization problems, using expert systems which offer a nonflexible way of dealing with the rapidly changing conditions of the real-world supply chain. The incorporation of machine learning models offers a comprehensive solution to address the limitations of current systems, presenting a more adaptive and holistic approach to tackling these challenges.

3.1 RESEARCH ABOUT CURRENT PROBLEMS IN THE SUPPLY CHAIN

Shipping cost is a critical factor in supply chain management and business operations, impacting various aspects of a company's performance, profitability, and competitive advantage. The Pharmaceutical shipping industry in particular globally faces a lot of challenges that include the proper working of the whole logistic process to deliver the goods from the source to their destination without incurring delays. Currently, the global supply chain management industry is in strife with issues such as delayed supply allocation due to backlogs and congestion in supply, drugs, and other time-sensitive entities that need to be delivered within certain time constraints. Certain disease outbreaks lead to a massive surge in demand for certain drugs which need to be manufactured and then transferred in the shortest duration. This designation ensures that crucial items or partners are readily available and receive immediate attention in the event of disruptions, such as shortages, delays, or emergencies. These components or suppliers are considered essential for the smooth functioning of the supply chain and are given special attention or priority in terms of procurement, production, and logistics.

In this research paper, the authors aim to address a specific facet of the supply chain challenge within the pharmaceutical industry-namely, the reduction of shipping costs. The approach taken involves predicting the most likely shipping cost based on historical order data. This approach holds promise for benefiting both consumers and sellers, as it provides an equitable and impartial estimate of shipping cost per unit weight, taking into account various factors such as order quantity, drug production costs, shipping expenses, and delivery schedules.

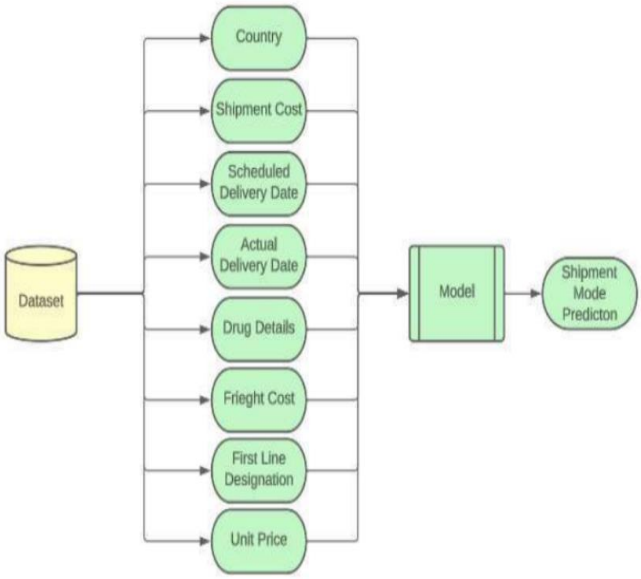

Fig -1: Model Overview

In a real-life supply chain, there exists certain factors which majorly influence a product's supply cost such as manufacturing charges and the rate of degradability of the product as well[. Events in the course of humanity such as ecological disasters and virus outbreaks increase the demand for the product in a certain localized region. The models that need to be developed would take into account the major factors that provide concrete reasoning for the selection of the delivery mode while being resistant to outliers and deviations that may arise due to erroneous data. Major factors to be considered should be the geographic region the drug would be shipped to, country of origin, expected and actual dates of delivery and dispatch of the drug, the importance of the drug denoted by its first-line designation, and the costs associated with creating the drug and the shipping costs

3.2 MODELS USED:

A. Decision Trees

Decision Trees manifest as a hierarchical, flowchart-like structure that begins with a root node and progressively splits data based on features. For our SCM scenario, this structure entails attributes like product characteristics, delivery schedules, and transportation costs. The primary objective is to construct a decision tree that accurately identifies the most optimal mode of transportation for pharmaceutical products. At each internal node, the Decision Tree algorithm employs criteria such as information gain or entropy to select the feature that best segregates the data. This process continues until leaf nodes are reached, representing the final recommendations for the ideal transportation mode.

One of the distinctive advantages of Decision Trees in the realm of SCM is their interpretability. The resulting tree structure can be easily understood and visualized, facilitating comprehension of the rationale behind transportation mode recommendations. However, it is important to note that Decision Trees can be prone to overfitting, particularly in the complex and dynamic SCM environment. To address this, techniques like pruning, depth constraints, or the utilization of ensemble methods such as random forests can be employed to ensure that the recommended transportation modes generalize effectively to new, unseen data In summary, within the context of pharmaceutical SCM, Decision Trees offer a valuable tool for predicting the optimal mode of transport, promoting efficiency, cost-effectiveness, and competitiveness within the supply chain.

B. Random Forest

Small-scale pharmaceutical companies face a complex challenge in optimizing their supply chain operations, which encompass multiple variables like demand fluctuations, inventory management, production planning, and transportation logistics. The Random Forest algorithm has emerged as a promising tool for handling these multifaceted prediction problems.

However, the Random Forest algorithm has limitations that need to be considered. One major constraint is its "black-box" nature, which means that the model's decision-making process is difficult to understand. While Random Forest can make accurate predictions, it lacks the transparency required for decision-makers to fully comprehend the rationale behind its recommendations. This opacity can pose challenges in highly regulated industries like pharmaceuticals, where decision-making often requires justification and adherence to specific guidelines.

Random Forest is also susceptible to overfitting, especially when applied to high-dimensional datasets or data with noise or outliers. Overfitting occurs when a model becomes too complex and learns the training data too well, leading to poor generalization of unseen data and potentially erroneous decision-making. Additionally, Random Forest may not perform optimally with imbalanced datasets, where one class or outcome is significantly more prevalent than others. In such cases, the algorithm may bias predictions towards the majority class, potentially leading to suboptimal decision outcomes, especially in scenarios where rare events or minority classes are of critical importance.

Finally, the effectiveness of Random Forest depends on the quality and quantity of data available. In situations where data is limited or noisy, the algorithm's predictive performance may be compromised.

C. XG Boost

XGBoost (Extreme Gradient Boosting) emerges as a potent machine learning algorithm, offering profound solutions to the intricate challenges encountered by small-scale pharmaceutical companies in managing their supply chains (SMC). By amalgamating ensemble learning with gradient boosting techniques, XGBoost constructs resilient predictive models adept at unraveling complex data relationships. This results in more precise forecasts and augments the decision-making process significantly.

Moreover, XGBoost addresses overfitting concerns by incorporating regularization techniques, ensuring the models generalize effectively to new data, thereby furnishing reliable predictions. Its capability to manage imbalanced data by assigning varying weights to classes is particularly advantageous in SMC scenarios, where rare events or less prevalent products hold pivotal significance.

In addition to its efficacy, XGBoost's efficiency and scalability are noteworthy attributes, enabling it to process extensive datasets in real-time, thereby enhancing operational efficacy in SMC decision-making. Furthermore, XGBoost enhances model interpretability by furnishing insights into feature importance, facilitating better-informed decisions and judicious resource allocation strategies.

Overall, XGBoost equips small-scale pharmaceutical companies with the tools necessary for data-driven decision-making, optimization of SMC operations, and enhancement of competitiveness within the pharmaceutical sector. Its robustness in handling complex data dynamics, coupled with its efficiency, scalability, and interpretability, renders XGBoost an invaluable asset in navigating the multifaceted landscape of supply chain management in the pharmaceutical industry. Through its capabilities, XGBoost empowers companies to not only adapt to the evolving demands of their supply chains but also to thrive amidst the challenges posed by an increasingly competitive market environment.

4. METHODOLOGY:

A. Dataset Used

This Dataset contains information about the pharmaceutical companies who are supplying medicines Internationally for curing various diseases such as Reveal G3 Rapid for HIV, Nevirapine for HIV, Lamivudine for Hepatitis B virus, Didanosine for HIV, artemether/lumefantrine for malaria Infections, etc. These medicines are of various brands such as Generic, Aluvia, Bioline, Videx, etc. and their dosage form is Test kit, Capsule, Tablet, Oral solution, Chewable/dispersible tablet, and Injection. These goods can be transported through various modes of transport such as air, Truck, Air Charter, and Ocean.

The supply chain dataset comprises 2909 orders and 19 attributes such as 'Sub Classification' (branch of medicine), 'Unit of Measure (Per Pack)' (packaging unit), 'Line item Quantity' (quantity to be transported), 'Line Item Value' (intended impression), 'Pack Price' (total cost), 'First Line Designation' (product preference), 'Weight (Kilograms)' (total weight), 'Freight Cost (USD)' (transportation cost), 'Unit Price' (single unit price), 'Line Item Insurance (USD)', 'Country' (origin), 'Shipment Mode' (transport mode), delivery dates (scheduled, received, and recorded), and corresponding years, months, and days.

	Shipment Mode	Sub Classification	Unit of Measure (Per Pack)	Line Item Quantity	Line Item Value	Pack Price	Unit Price	First Line Designation	Weight (Kilograms)	Line Item Insurance (USD)
count	2909.000000	2909.000000	2909.000000	2909.000000	2.909000e+03	2909.000000	2909.000000	2909.0	2909.000000	2909.000000
mean	0.197319	2.040664	80.936404	15800.745961	1.585401e+05	37.947477	1.240034	0.0	2058.106222	76.694190
std	0.592462	1.351195	95.275116	36420.694323	2.386844e+05	61.402235	3.800968	0.0	4390.126206	59.362311
min	0.000000	0.000000	1.000000	1.000000	0.000000e+00	0.000000	0.000000	0.0	0.000000	0.000000
25%	0.000000	1.000000	30.000000	500.000000	1.046700e+04	6.650000	0.110000	0.0	140.000000	15.250000
50%	0.000000	2.000000	60.000000	2906.000000	6.435700e+04	23.000000	0.390000	0.0	904.000000	65.210000
75%	0.000000	2.000000	100.000000	12315.000000	2.236800e+05	70.000000	0.890000	0.0	2517.000000	135.292755
max	3.000000	5.000000	1000.000000	515000.000000	2.801262e+06	1250.000000	30.000000	0.0	154780.000000	199.910000

	Scheduled Delivery Date Year	Scheduled Delivery Date Month	Scheduled Delivery Date Day	Delivered to Client Date Year	Delivered to Client Date Month	Delivered to Client Date Day	Delivery Recorded Date Year	Delivery Recorded Date Month	Delivery Recorded Date Day
count	2909.000000	2909.000000	2909.000000	2909.000000	2909.000000	2909.000000	2909.000000	2909.000000	2909.000000
mean	2011.552767	6.644208	15.520110	2011.554486	6.649708	17.512891	2011.558267	6.654864	15.637332
std	2.110525	3.406784	8.721546	2.111947	3.407757	8.799516	2.116076	3.412569	8.790164
min	2006.000000	1.000000	1.000000	2006.000000	1.000000	2.000000	2006.000000	1.000000	1.000000
25%	2010.000000	4.000000	8.000000	2010.000000	4.000000	10.000000	2010.000000	4.000000	8.000000
50%	2012.000000	7.000000	15.000000	2012.000000	7.000000	18.000000	2012.000000	7.000000	16.000000
75%	2013.000000	10.000000	23.000000	2013.000000	10.000000	25.000000	2013.000000	10.000000	23.000000
max	2015.000000	12.000000	31.000000	2015.000000	12.000000	33.000000	2015.000000	12.000000	31.000000

Fig -2: Summary of Data Set

	Unit of Measure (Per Pack)	Line Item Quantity	Line Item Value	Pack Price	Unit Price	Line Item Insurance (USD)
count	10324.000000	10324.000000	1.032400e+04	10324.000000	10324.000000	10037.000000
mean	77.990895	18332.534870	1.576506e+05	21.910241	0.611701	240.117626
std	76.579764	40035.302961	3.452921e+05	45.609223	3.275808	500.190568
min	1.000000	1.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	30.000000	408.000000	4.314593e+03	4.120000	0.080000	6.510000
50%	60.000000	3000.000000	3.047147e+04	9.300000	0.160000	47.040000
75%	90.000000	17039.750000	1.664471e+05	23.592500	0.470000	252.400000
max	1000.000000	619999.000000	5.951990e+06	1345.640000	238.650000	7708.440000

Fig -3: Description of Data Set

B. Data Cleaning

Data cleaning involves rectifying or eliminating inaccurate, corrupted, wrongly formatted, duplicate, or incomplete data within a dataset. When amalgamating multiple data sources, there are numerous opportunities for data replication or mislabeling. If data is flawed, the outcomes and algorithms become unreliable, despite appearing correct. To ensure the highest quality data, certain criteria such as validity, accuracy, completeness, consistency, and uniformity must be met within the dataset.

In the given dataset, several data cleaning operations were performed, including the removal of unnecessary columns, elimination of outliers and NaN values, and rectification of erroneous string values in various features.

1) Elimination of redundant columns:

Columns such as 'PQ #', 'PO / SO #', 'ASN/DN #', 'Managed By', 'Fulfill Via', 'Vendor INCOTerm', 'PQ First Sent to Client Date', 'Product Group', 'Vendor', and 'Ï»¿ID' were removed as they do not contribute to the objective or are irrelevant for predicting shipment mode.

2) Removal of outliers and NaN values:

Certain features like Line Item Insurance, Dosage, and Shipment mode contained NaN values, which were replaced with median values where applicable. Outliers were also eliminated using the IQR method to ensure data consistency.

3) Rectification of erroneous string values:

In features like 'Freight Cost (USD)', 'Weight (Kilograms)', and 'PO Sent to Vendor Date', which are supposed to be of integer type, strings like 'Freight Included in Commodity Cost', 'Invoiced Separately', 'See', 'Weight Captured Separately', 'Date Not Captured', and 'N/A - From RDC' were removed along with the corresponding rows containing these strings.

Feature Engineering:

Feature engineering is utilized on the dataset to refine the data and extract characteristics suitable for the learning process. As part of this process, feature engineering is applied to 'Scheduled Delivery Date', 'Delivered to Client Date', and 'Delivery Recorded Date' to partition them into the distinct year, month, and date components, thereby refining the data and extracting features conducive to the learning process.

Shipment Mode	Scheduled Delivery Date Year	Scheduled Delivery Date Month	Scheduled Delivery Date Day	Delivered to Client Date Year	Delivered to Client Date Month	Delivered to Client Date Day	Delivery Recorded Date Year	Delivery Recorded Date Month	Delivery Recorded Date Year
0	2007	2	27	2007	2	29	2007	2	2007
0	2007	6	6	2007	6	8	2007	6	2007
0	2007	6	19	2007	6	21	2007	6	2007
0	2007	6	19	2007	6	21	2007	6	2007
0	2007	10	15	2007	10	17	2007	10	2007
...	...	...	...	...	...	...	...	...	...
2	2011	3	4	2011	3	5	2011	3	2011
2	2010	12	20	2011	1	13	2011	1	2011
0	2012	10	8	2012	10	10	2012	10	2012
0	2010	9	7	2010	11	19	2010	11	2010
0	2013	4	11	2013	4	13	2013	4	2013

Fig -4: Feature Engineering Performed on data features

Data Preprocessing:

In the process of data preprocessing, a crucial step involves converting categorical values into numerical representations to facilitate their utilization within machine learning models, which typically operate exclusively on numerical data. This conversion is achieved through label encoding, where each unique categorical value is assigned a specific numerical code. In the case of the 'Shipment Mode' feature, label encoding is applied to ensure compatibility with machine learning algorithms. Specifically, the categorical values such as 'Air', 'Air Charter', 'Ocean', and 'Truck' are encoded as follows:

- Air: 0

- Air Charter: 1

- Ocean: 2

- Truck: 3

By transforming categorical data into numerical equivalents, it becomes feasible to effectively integrate these features into machine learning models for further analysis and predictive tasks.

C. Data Visualisation

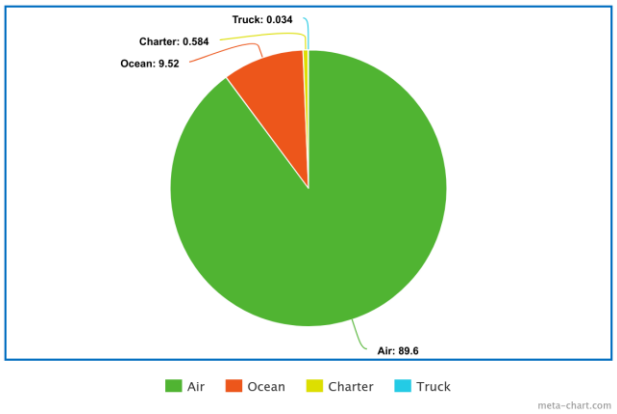

Fig 5: Pie Chart Describing the Shipment Mode.

Analysis of the pie chart (Figure 5) reveals that the predominant shipment mode is air, constituting approximately 89.9% of the total. Following air transportation, ocean freight accounts for around 9.52% of the shipments. Subsequently, Air Charter emerges as the third most utilized mode, representing approximately 0.584%, while Truck transportation comprises a minimal 0.0334% of the total shipments.

D. Architecture.

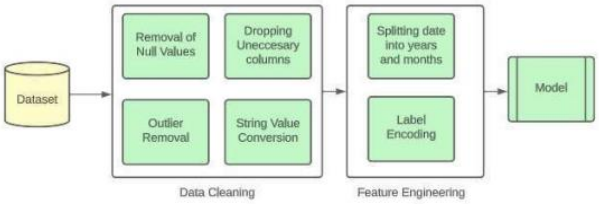

Fig 6: Model Architecture.

The provided dataset necessitates preprocessing to eliminate irrelevant data features, and data augmentation can further enhance the accuracy of machine learning (ML) models. Initially, the data cleaning phase involves dropping non-outliers using the interquartile range method. Additionally, NaN values in string data are addressed. Data augmentation entails feature engineering on the shipment delivery column, where the date, initially in string format, is dissected into its constituent components of day, month, and year.

Upon completion of feature engineering, the preprocessed data is fed into the model. The dataset is utilized to train models such as the Decision Tree Classifier, Random Forest Classifier, and XGBoost. Particularly, XGBoost constructs an ensemble of decision trees sequentially, aiming to rectify errors made by preceding trees. It employs gradient descent optimization for efficient training, reducing training time while upholding high accuracy. Decision trees offer interpretability and visualizability due to their transparent decision-making logic.

In contrast, Random Forests generate multiple decision trees during training. Each tree is trained on a randomly chosen subset of the data and features, introducing variability among trees and mitigating the risk of overfitting. This approach ensures that each tree perceives the data differently, enhancing the model's robustness.

5. RESULTS AND DISCUSSION:

The comparison between the Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and XGBoost Classifier (XGBC) was conducted based on several key metrics, including accuracy, scalability, speed, and robustness.

1. Accuracy and Scalability:

In terms of accuracy and scalability, it was observed that the XGBoost Classifier (XGBC) outperformed both the Decision Tree Classifier (DTC) and Random Forest Classifier (RFC). As the size of the dataset increased, the accuracy of the classifiers also improved.

Specifically, for a dataset with a larger number of instances, the XGBoost Classifier achieved the highest accuracy among the three classifiers. The Decision Tree Classifier attained an accuracy of 80.8571%, the Random Forest Classifier achieved an accuracy of 90.0481%, and notably, the XGBoost Classifier demonstrated the highest accuracy of 95.4199%.

This observation suggests that the XGBoost Classifier is particularly effective in handling larger datasets, leading to improved accuracy compared to the other classifiers.

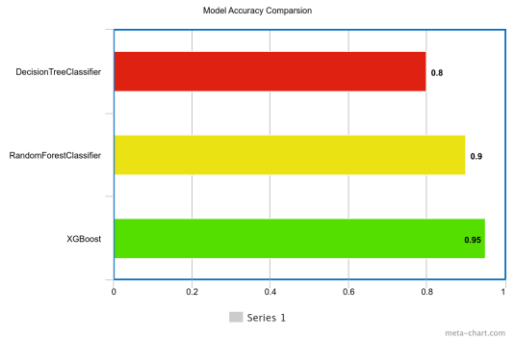

Fig7: Comparison between Accuracy of Different ML Model.

Based on the information provided in Figure 7, it is evident that the accuracy of three algorithms was compared, and notably, the XGBoost classifier exhibited the highest accuracy among them.

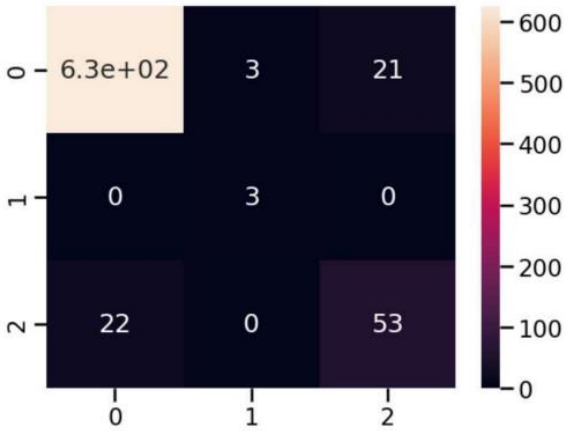

Fig 8: Confusion Matrix of XGBoost Classifier Algorithm.

The analysis of the confusion matrix (Figure 8) indicates that the model is effectively predicting correct values based on the training data. This conclusion is drawn from observing that the diagonal elements of the confusion matrix, corresponding to True Positive and True Negative values, are notably higher compared to the non-diagonal elements, representing False Positive and False Negative values. The relatively smaller values of False Positive and False Negative instances further support the notion that the model is performing well in accurately classifying the data, thus demonstrating its effectiveness in training.

	precision	recall	f1-score	support
0	0.99	0.97	0.98	531
2	0.74	0.88	0.80	51
accuracy			0.96	582
macro avg	0.86	0.93	0.89	582
weighted avg	0.97	0.96	0.96	582

Fig 9: Classification Report of XGBoost Classification Algorithm.

The analysis of the classification report (Figure 9) clearly demonstrates the strong performance of the XGBoost Classifier algorithm model. With an impressive accuracy of 96.4199%, coupled with other metrics such as F1-Score and recall close to 1, it is evident that the model has excelled across various evaluation criteria.

Moreover, the reported hyperparameters utilized in the model further underline its effectiveness. Parameters such as 'subsample: 0.5', 'num_class: 10', 'n_estimators: 1100', 'max_depth: 10', 'learning_rate: 0.1', and 'colsample_bytree: 0.7999' reflect a well-tailored configuration that has contributed to the high performance observed in the classification report.

1. Speed and Robustness

In terms of speed and robustness, the models exhibited varying performance metrics. The Decision Tree Classifier (DTC) required approximately 0.3 seconds to build the model, while the Random Forest Classifier (RFC) took around 0.6 seconds to both build and test the model on the test split. Conversely, the XGBoost Classifier (XGBC) took approximately 0.94 seconds to build the model and zero seconds to test the model on the test split.

The time complexity for building each model varies. For Decision Tree Classification, it is O(nmlog(m)), potentially reaching O(n^2 m) in the worst-case scenario. For RFC, the time complexity is O(n_trees * n * m * log(m)). XGBoost Classifier's time complexity is linear with respect to the number of boosting rounds, with the complexity of building each tree similar to that of a single decision tree (O(n * m * log(m))).

These observations highlight notable differences in the computational efficiency and robustness of the models, with XGBoost Classifier demonstrating both fast build time and robust performance.

6. CONCLUSION

In conclusion, this research aimed to identify crucial features for training a machine learning (ML) model utilized in the pharmaceutical supply chain industry to predict the optimal weight prediction. Through extensive exploratory data analysis, data cleaning, and feature extraction processes, insights were gleaned from the dataset. Subsequently, various ML models were trained and evaluated across multiple performance metrics, with Decision Trees (DT), Random Forests, and XGBoost emerging as the top-performing models.

It was observed that tree-based ML models exhibited optimal performance compared to other standard ML algorithms, such as regression. However, the choice of algorithm depends on the specific task at hand. Notably, the XGBoost Classifier algorithm yielded the most accurate classifier models, achieving an impressive accuracy of 96.4199%. Furthermore, while the Decision Tree Classifier displayed the highest residual error, Random Forest exhibited a lower residual error, with the XGBoost Classifier showcasing the least residual error.

For future work, it is imperative to explore further enhancements to model performance and explore additional features that may contribute to improving prediction accuracy. Additionally, investigating the applicability of other advanced ML techniques and ensemble methods could provide valuable insights for optimizing the prediction of optimal shipment modes in the pharmaceutical supply chain industry.

ACKNOWLEDGEMENT

Our special thanks to my research guide Mr. Ajinkya Valanjoo and Mrs. Kusum Kadam for his continuous motivation during this research. I express my special thanks to my co-authors for helping me during the preparation of this research paper. I also thank all the experts who have contributed to the development of this research topic.

REFERENCES

1. Detwal, P. K., Soni, G., Jakhar, S. K., & Kayikci, Y. (2023). Supply chain fraud prediction based on XGBoost method. Journal of Business Research.

2. Ni, S. F., Peng, Y., Peng, K., & Liu, Z. J. (2022). Supply chain demand forecasting based on the SSA-XGBoost model. Journal of Computer and Communications, 10(1), 71-83.

3. Dua, S., Singh, A. P., & Kaur, S. (2022). XGBoost with Q-learning for complex data processing in business logistics management. Expert Systems with Applications, 194, 116641.

4. Tirkolaee, M. B., Soleymani, S., & Ahmadi, Z. (2021). A novel ensemble model based on XGBoost and hybrid features for predicting drug-induced liver injury. Expert Systems with Applications, 180, 115334

5. Malik, A., Hassan, S. H., & Islam, R. (2022). Exploring the impact of supply chain disruptions on production planning and control using a XGBoost-based approach. Journal of Manufacturing Systems, 106, 100588.

6. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

7. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1266-1284.

8. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

9. Ribeiro, M. T. (2019). Explainable artificial intelligence (XAI): A survey. arXiv preprint arXiv: 1910.10023.

10. A Decision Tree-Based Approach for Selecting the Optimal Transportation Mode for Pharmaceutical Products, by Sharma, S., & Kumar, A. (2019).

11. Using Decision Trees to Analyze the Impact of Transportation Mode on Pharmaceutical Supply Chain Performance by Wang, X., & Zhang, L. (2015).

12. A Review of Supply Chain Management Models for Pharmaceutical Products by Sharma, S., & Kumar, A. (2020).

13. Factors Affecting the Supply Cost of Pharmaceutical Products: A Case Study by Chen, J., & Wang, S. (2019).

14. Guiffrida, Alfred L. "Recent trends in supply chain delivery models." International Journal of Industrial and Manufacturing Engineering 8.6 (2014): 1823-1826.

15. A review of machine learning applications in supply chain management" Boudia, Moursli, and Ait-Kadi Computers & Industrial Engineering