Predicting Pharmaceutical Product Discontinuation Using Decision Tree and Random Forest Algorithms Based on Product Attributes

Susilo Hartono[1,*], Nur Azizah[2]

[1]*Information Systems and Technology, Muhammadiyah University of Pringsewu Lampung, Indonesia*

(Received: December 20, 2024; Revised: January 25, 2025; Accepted: April 20, 2025; Available online: July 23, 2025)

Abstract

This study aims to predict the discontinuation of pharmaceutical products using machine learning models, focusing on key product attributes such as manufacturer, composition, price, and packaging. A comprehensive dataset of over 250,000 pharmaceutical products from India was analyzed, with two models—Decision Tree and Random Forest—being employed for prediction. The models were evaluated based on accuracy, precision, recall, and F1-score. The Random Forest model outperformed the Decision Tree with a higher accuracy, but both models struggled with the imbalanced dataset, showing low recall for the minority class (discontinued products). Feature importance analysis identified manufacturer and composition as the most influential factors in predicting product discontinuation. These findings offer valuable insights for pharmaceutical companies in managing product portfolios and optimizing their lifecycle strategies. Despite limitations in data quality and class imbalance, this study provides a foundation for future research, suggesting the integration of additional data sources and the application of deep learning techniques to further enhance prediction accuracy.

Keywords: Decision Tree, Discontinuation Prediction, Machine Learning, Pharmaceutical Products, Random Forest

1. Introduction

The pharmaceutical industry is a pivotal segment of the global economy, significantly influencing public health and healthcare accessibility worldwide. It comprises a complex network of various stakeholders, including manufacturers, regulators, healthcare providers, and patients, all contributing to the overall functioning of this multifaceted sector. The industry's primary objective lies in the research, development, production, and marketing of medications that aim to improve health outcomes and quality of life. However, the industry faces numerous challenges, including regulatory scrutiny, patent expirations, market competition, and constant pressure for innovation. These factors can significantly impact dynamics within the pharmaceutical market, particularly regarding product discontinuation.

Product discontinuation within the pharmaceutical industry can be attributed to several determinants, including declining Research and Development (R&D) success rates, patent expirations, regulatory challenges, market competition, and shifts in public health needs. Recently, the pharmaceutical sector has indeed been scrutinized for not innovating rapidly enough to meet the growing demands of global health issues [1]. Market conditions reflect this trend; companies experience pressures from generic competition and stringent regulatory frameworks designed to ensure drug safety and efficacy. Product discontinuation may thus be viewed as a means through which companies rationalize their pipelines, focusing investment on the most promising molecules while discarding less viable products [2], [3].

The ramifications of product discontinuation are multifaceted and can lead to significant consequences for both the companies involved and the healthcare system at large. Discontinuation of a product, especially one addressing unmet medical needs, can exacerbate health disparities, particularly in developing nations where access to innovative medicines is already limited [1]. Pharmaceutical companies must navigate these challenges carefully, balancing the need for robust pipelines with potential repercussions on public health. Moreover, the market often requires that

***Corresponding author: Susilo Hartono (susilohartono@umpri.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v5i2.101

companies adopt new strategies to extend market exclusivity and derive returns from existing portfolios, leading to strategic shifts such as mergers and acquisitions [4], [5]. These strategic frameworks enable firms to consolidate resources, address patent cliffs, and maximize yield from their R&D investments while sustaining levels of innovation and operational efficiency.

Predicting the discontinuation of pharmaceutical products presents a significant challenge within the industry that stems from a multitude of complex factors. The dynamic nature of the pharmaceutical market, characterized by rapid advancements in technology, evolving regulatory landscapes, and shifting public health perspectives, complicates the forecasting process surrounding product viability and longevity. Discontinuation may occur due to various reasons, ranging from clinical efficacy and safety concerns to financial and strategic considerations. The challenge lies not only in identifying the indicators of potential discontinuation but also in accurately interpreting them amidst the myriad external forces influencing product trajectories.

One primary factor that complicates predictions related to product discontinuation is the intricacy involved in the clinical trial process. For instance, a substantial percentage of clinical trials experience early termination, with existing research showing that approximately 30% of trials in certain medical specialties do not complete as planned [6]. The reasons for discontinuation can vary widely, including inadequate recruitment, poor study design, and unforeseen adverse events [7]. This diversity in causes indicates the unpredictable nature of clinical trial outcomes, creating barriers to accurate forecasting of whether a product may be withdrawn from the market post-approval or during the developmental stages. Furthermore, inconsistent reporting practices regarding the rationale for trial discontinuation further obscure understanding of trends and raise questions about the factors influencing these decisions [6], [7].

Machine Learning (ML) technologies have emerged as powerful tools to enhance predictive capabilities within clinical settings, including the pharmaceutical industry. Recent studies have demonstrated that ML models can forecast the likelihood of drug discontinuation with considerable accuracy. For example, a Generalized Linear Model (GLM) provided an 80% accuracy rate in predicting treatment discontinuation specifically in patients receiving biologic treatments for psoriasis, effectively identifying causative factors such as loss of efficacy and adverse reactions [8]. While promising, the integration of ML depends heavily on the quality and availability of historical data. Inconsistent data collection and reporting practices can lead to significant biases and inaccuracies, complicating predictions about product continuity [9].

The development of predictive models for pharmaceutical product discontinuation is increasingly vital given the various challenges the industry faces in ensuring product viability. The research delves into understanding the factors that influence discontinuation decisions and underscores the necessity for robust predictive frameworks that can account for these diverse influences. This study aligns with broader trends in the pharmaceutical sector emphasizing data-driven decision-making and adaptive strategies based on predictive analytics.

A central focus of this study is leveraging advanced computational models to enhance forecasting capabilities. Deep learning methodologies, which have shown promise in various domains including forecasting drug dissolution profiles and predicting pharmacological outcomes, can be invaluable in addressing the complexities surrounding product discontinuation [10]. For instance, Artificial Neural Networks (ANNs) have demonstrated significant effectiveness in predicting outcomes from pharmaceutical formulations, highlighting their potential for application in predicting product lifecycle challenges [10]. These methodologies allow for the identification of potential indicators of discontinuation and facilitate the integration of diverse datasets reflecting clinical, economic, and regulatory influences.

The scope of this research is centered around a comprehensive dataset that includes over 250,000 pharmaceutical products available in India. This dataset provides crucial details such as the product's brand name, active ingredients, manufacturer, packaging, and other relevant attributes. The primary focus is on attributes like active ingredients, which are essential in understanding the therapeutic value of the drug, manufacturer, which may reflect market competition, and packaging, which often correlates with consumer preferences and market trends. These features serve as key indicators that could help predict whether a product will be discontinued in the market. Additionally, the dataset is structured to facilitate the analysis of drug pricing trends, product availability, and market behavior, offering valuable insights for predicting product lifecycle outcomes. The central research question for this study is how can machine learning models predict pharmaceutical product discontinuation based on available features? This question is addressed

by utilizing machine learning techniques, particularly classification models like Decision Trees and Random Forest, to analyze the relationships between the dataset's features and the likelihood of a product being discontinued. By exploring these patterns, the research aims to identify significant predictors of product discontinuation, helping pharmaceutical companies to make informed decisions about product portfolio management. The predictive models developed in this study can serve as tools for forecasting product lifecycle events, contributing to more strategic decision-making in the pharmaceutical industry.

2. Literature Review

2.1. Previous Studies on Pharmaceutical Product Discontinuation

The exploration of pharmaceutical product discontinuation has garnered attention from various studies aiming to predict outcomes related to the viability and lifecycle of medications. Previous research has demonstrated that numerous factors contribute to the discontinuation of pharmaceutical products, ranging from clinical efficacy to strategic business decisions. This overview synthesizes findings from key studies that sought to understand and predict pharmaceutical product discontinuation or similar outcomes. One significant study by Weygandt et al examined the discontinuation and nonpublication of clinical trials in the context of pharmacologic treatments specifically for posttraumatic stress disorder among military veterans. This work uncovers the multifactorial nature of clinical trial discontinuation, pointing out that ClinicalTrials.gov often lists a single reason for discontinuation, potentially oversimplifying the complexity involved. Limitations of this study included a small sample size and a focus on a niche population, which restricted generalizability. Nonetheless, it established a precedent for how clinical trial dropout rates can inform broader patterns of product discontinuation [11].

Another noteworthy investigation by Zeng et al conducted a meta-analysis of abandoned cardiovascular, renal, and metabolic therapeutics over a decade (2011-2022). This research highlighted the impact of known side effects on discontinuation decisions, emphasizing that strategic reasons (e.g., budget constraints and shifting market priorities) often override safety or efficacy considerations. The findings indicate that pharmaceutical companies frequently in-license drugs that were previously discontinued, thus suggesting ongoing interest in potentially viable compounds despite past failures [12]. This work underscores the need for predictive models that could evaluate changing market conditions and aggregate historical data on discontinued products to derive insights about future viability. Hafner et al addressed pharmaceutical systems strengthening and the life cycles of pharmaceutical products from production to use. Their findings elaborate on how understanding the pharmaceutical ecosystem—including factors that precipitate discontinuation—can inform systems-level approaches to measurement and evaluation. They identify the significance of pharmaceutical innovation at every stage, from research and development through to market entry and utilization. This systems-oriented perspective provides a foundation for constructing predictive models by fostering an understanding of the entire lifecycle of pharmaceutical products and highlighting factors conducive to both success and failure [13].

Pitchayajittipong et al focused on the pharmaceutical production landscape within Thai hospitals, revealing that supply inadequacies could lead to the production of nonsterile preparations. Their study indicated that the absence of commercially available products in necessary dosage forms necessitates locally produced alternatives. This finding aligns with the ongoing debate about the balance between drug availability and the potential discontinuation of products that fail to meet specific health needs or regulatory standards [14]. The qualitative study by Conder et al explored challenges associated with pharmaceutical drying processes, emphasizing operational factors that can lead to product stability issues and ultimately influence their market viability [15]. Insights from their findings can be critical for developing predictive measures concerning the sustainability and durability of drug formulations.

2.2. Data Mining Techniques in Healthcare

The application of data mining techniques in healthcare has become increasingly important as the demand for effective patient care and the utilization of vast amounts of medical data expand. Various studies have explored different algorithms and methodologies, particularly focusing on ML algorithms such as Decision Trees and Random Forests. This review addresses previous works that applied these and other data mining techniques for various healthcare-related objectives, reflecting on their significance, challenges, and future directions. In their exploration of the role of data

mining in healthcare, Alzahrani and Safhi emphasized the necessity of leveraging these techniques to manage big data effectively. Their research articulated how healthcare administrators recognized the critical trends associated with data mining, which have contributed significantly to improving patient services and outcomes [16]. The integration of data mining, particularly machine learning methods, has enabled healthcare facilities to analyze large datasets more effectively, leading to enhanced decision-making processes.

Pika et al focused on the application of privacy-preserving process mining techniques, illustrating the challenges associated with preserving privacy while utilizing healthcare data. Their work provided insights into process mining algorithms frequently used in healthcare settings, emphasizing the need for accuracy while managing data privacy concerns [17]. This underscores the critical need for healthcare organizations to implement robust analytical frameworks that respect patient privacy while deriving meaningful insights from data. Oliveira et al provided an integrative review of various data mining techniques and their applications in healthcare for generating knowledge related to patient care. Their analysis presented an overview of the benefits and challenges faced in the integration of data mining into clinical processes. They discussed the inherent ethical considerations involved, which is pivotal in navigating the complexities of using patient data for predictive analytics in clinical settings [18].

Diriba's work on developing a risk level prediction model for cardiovascular diseases illustrated the practical applications of data mining techniques for clinical Decision Support Systems (DSS) in Ethiopia. By employing algorithms like C4.5, Diriba demonstrated that data mining technologies could be instrumental in forecasting patient needs, thus facilitating timely and appropriate treatment decisions [19]. The ability to predict patient conditions emphasizes how the implementation of machine learning can enhance clinical practices significantly. Research by Zawad investigated the emerging trends of data mining within the healthcare context of Bangladesh. The study indicated that data mining has facilitated the identification and detection of various diseases, as well as the development of policy recommendations tailored to healthcare needs. By highlighting the challenges faced in a traditional healthcare system, Zawad underscored the transformative potential of data mining methods to support healthcare research and strategy developments globally [20].

2.3. Algorithms for Classification

Classification techniques are pivotal in machine learning and data mining, particularly in healthcare. A Decision Tree is a flowchart-like structure where each internal node represents a feature (attribute), each branch represents a decision rule, and each leaf node represents an outcome (class label). The path from the root to a leaf signifies the decision-making process based on rules derived from the data, with algorithms like Classification and Regression Trees (CART) commonly used to construct these trees. One key advantage of Decision Trees is their interpretability, as they provide clear and transparent reasoning behind predictions, making them suitable for applications where understanding the decision-making process is crucial [21]. Additionally, they do not assume a specific data distribution, allowing them to handle diverse datasets without requiring pre-defined structures [22]. Furthermore, Decision Trees can manage both numerical and categorical data without significant preprocessing [23]. However, Decision Trees also have limitations, such as overfitting, especially with deep trees that may model noise rather than meaningful patterns [24]. They also exhibit instability, where slight changes in the data can lead to significantly different splits, affecting predictive accuracy [21]. Additionally, they may show bias towards dominant classes, particularly in imbalanced datasets [25].

Random Forest builds upon the Decision Tree model by constructing multiple decision trees during training and outputting the mode of the classes (for classification) or the mean prediction (for regression) from individual trees [21]. This ensemble technique utilizes bagging (bootstrap aggregating) to reduce variance and improve model accuracy. One key advantage of Random Forest is its robustness against overfitting, as it aggregates the results from multiple trees, making it less likely to overfit compared to a single Decision Tree [22]. Additionally, Random Forest is resistant to missing values and outliers, maintaining accuracy even when data is incomplete or contains anomalies [21]. It also provides insights into feature importance, helping practitioners understand the variables that drive predictions [23]. However, the model has limitations, such as interpretation difficulty due to the complexity of the ensemble method, which makes it harder to explain decisions to stakeholders compared to individual Decision Trees [25]. Furthermore, Random Forest can lead to increased computation time and large memory requirements, especially when constructing multiple trees with large datasets, making it less suitable for real-time applications [26], [8]. Both Decision Trees and

Random Forests have established themselves in various healthcare applications, particularly in predicting outcomes, diagnosing diseases, and enhancing clinical decision-making. For instance, a study by Emam et al successfully employed machine learning techniques, including Random Forest, to predict treatment outcomes in psoriasis patients [8]. Their results emphasized the efficacy of these algorithms in improving diagnostic accuracy. Another study by Yang et al demonstrated the predictive capabilities of Random Forest in diagnosing SARS-CoV-2 using laboratory blood tests [27]. The model's adaptability across different patient populations highlights its robustness in clinical applications. Additionally, Liu et al analyzed the Random Forest model's efficacy in predicting the outcomes of chemoradiotherapy for patients with advanced cervical cancer, reinforcing its clinical relevance [23].

2.4. Relevant Formulas

The mathematical foundations of classification algorithms, particularly Decision Trees and Random Forests, are crucial for understanding their operational mechanics and effectiveness in various applications, including healthcare and data analytics. This overview focuses on key concepts such as Information Gain and the Gini Index, which are instrumental in Decision Trees, and discusses their significance in Random Forest implementations as well. Information Gain (IG) is a metric used to choose the ideal feature (attribute) to split the dataset at each node in a Decision Tree. It calculates how much information a feature provides about the class label. The essence of the concept is rooted in information theory, notably in the work of Shannon. The formula for calculating Information Gain is:

$$\mathrm{IG}(D, A) = H(D) - H(D|A) \tag{1}$$

Where H(D) is the entropy of the dataset (D) before the split and H(D|A) is the entropy of the dataset after the split on attribute (A) [28]. In this context, entropy measures the impurity or randomness in the class distribution of the dataset. A lower value of entropy indicates a purer node, leading to improved predictive accuracy.

The Gini Index is another popular criterion for evaluating a split's effectiveness. It is used in building Decision Trees (especially with the CART algorithm) and represents the probability of a randomly chosen element being incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. The Gini impurity can be computed as follows:

$$Gini(D) = 1 - \sum i = 1^{n} pi^{2} \tag{2}$$

Where pi is the proportion of instances belonging to class (i) and n is the number of classes within the dataset [21]. A lower Gini index indicates a more homogeneous node, thus guiding the selection process in tree construction.

3. Methodology

3.1. Exploratory Data Analysis (EDA) and Preprocessing

EDA is a critical step to understand the structure of the dataset and prepare it for machine learning models. The initial phase involves loading the dataset and performing a thorough review of its structure. The dataset, `indian_pharmaceutical_products_clean.csv` from Kaggle, contains various product attributes, including brand names, active ingredients, manufacturers, packaging details, prices, and the target variable `is_discontinued`, indicating whether a pharmaceutical product has been discontinued or is still available in the market.

The first task during EDA is to inspect and handle missing values in the dataset. Missing data is a common challenge in real-world datasets, and it can introduce biases or distort the results if not managed properly. In this research, we handle missing values by using the SimpleImputer class from `sklearn`. For numerical columns, such as price and quantities, we use the `mean` imputation strategy, where missing values are replaced by the mean of the respective column. For categorical columns, such as the `manufacturer_name` or `dosage_form`, we use the `most_frequent` imputation strategy, which replaces missing values with the most common value in the column. This approach ensures that no data is lost during preprocessing, maintaining the integrity of the dataset for analysis.

The next step in preprocessing involves feature engineering to enhance the dataset's usability. One notable transformation is the handling of the `active_ingredients` column, which often contains complex, nested data. The column's values are evaluated using a custom function, `parse_active_ingredients`, to parse and clean the ingredient

information. If the column contains lists in string format, the function extracts the names of the active ingredients and concatenates them into a single string. This new feature, `all_ingredients`, provides a cleaner, more usable format for analysis and modeling. Additionally, columns that are deemed redundant or too complex to parse are dropped. These include `product_id`, `brand_name`, and `active_ingredients`, which are not critical for predicting discontinuation. This cleaning process reduces noise and ensures that only the most relevant data remains.

3.2. Data Visualization

Visualization is a key component of EDA as it provides insights into the distribution and relationships between different variables in the dataset. In this research, we generate multiple visualizations to better understand the target variable `is_discontinued` and its relationship with other product attributes.

The first visualization is a count plot of the `is_discontinued` variable, which shows the distribution of products that are discontinued versus those that are still active. This plot helps to visualize the imbalance between the two classes and can serve as a starting point for further analysis. Following this, several other visualizations are created to explore the impact of different features on product discontinuation. For example, we generate histograms and kernel density estimation (KDE) plots to analyze the price distribution of pharmaceutical products. This helps identify any significant price variations that might correlate with a product's likelihood of discontinuation.

We also visualize the relationship between manufacturer and product discontinuation using a count plot grouped by `manufacturer`. This plot highlights the top manufacturers and their associated discontinuation rates, providing insights into which companies have more products discontinued than others. Similarly, we explore the relationship between therapeutic class and discontinuation using another count plot, which helps to identify if certain therapeutic categories are more likely to experience product discontinuation. These visualizations are not only useful for EDA but also provide intuitive insights that inform the feature selection process and model evaluation.

3.3. Feature Encoding and Scaling

Once the dataset has been cleaned and visualized, the next step is to encode categorical variables and scale numerical features to prepare them for machine learning algorithms. Label Encoding is used to transform categorical variables such as `manufacturer`, `dosage_form`, and `therapeutic_class` into numerical representations. This is necessary because most machine learning algorithms require numerical input, and encoding categorical data allows the model to process these features effectively.

After encoding categorical variables, we proceed to feature scaling using the StandardScaler from `sklearn`. Standardization is particularly important for many machine learning models that are sensitive to the scale of the input data. Although tree-based algorithms like Decision Trees and Random Forests are generally less sensitive to feature scaling, it is still a good practice to scale the data to ensure uniformity across the dataset. The StandardScaler is applied to all numerical features, transforming them into a distribution with a mean of zero and a standard deviation of one. This step ensures that all features contribute equally to the model training process.

3.4. Model Training

With the data prepared, we proceed to train the machine learning models. In this research, we use two popular classification algorithms: Decision Tree and Random Forest. These models are well-suited for the task of predicting whether a pharmaceutical product will be discontinued, as they can handle both categorical and numerical data effectively.

Before training, the dataset is split into training and testing sets using the train_test_split function from `sklearn`. The data is split in a 70-30 ratio, with 70% of the data used for training and the remaining 30% used for testing. This split is important for evaluating model performance and ensuring that the model can generalize well to unseen data. Additionally, the data is stratified based on the target variable `is_discontinued` to ensure that both classes (discontinued and active) are represented proportionally in both the training and testing sets.

The Decision Tree model is trained using the default settings, with Gini Impurity as the criterion for splitting nodes. Gini Impurity is a measure of how often a randomly selected element would be incorrectly classified. This algorithm is intuitive and interpretable, making it a good choice for initial modeling. The Random Forest model is an ensemble

method that constructs multiple decision trees during training and aggregates their predictions. We use 100 estimators (trees) for the Random Forest model, and the `n_jobs=-1` parameter is specified to utilize all available cores for faster training. This is particularly helpful when training on large datasets, as it speeds up the process by running multiple trees in parallel.

3.5. Evaluation and Model Comparison

After training both models, we evaluate their performance using common classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model, while precision and recall provide insights into the performance of the model with respect to the positive class (i.e., discontinued products). The F1-score is the harmonic mean of precision and recall, offering a balanced measure that takes both false positives and false negatives into account.

In addition to these metrics, we evaluate the models using confusion matrices, which provide a more detailed view of the model's performance by showing the true positives, true negatives, false positives, and false negatives. Confusion matrices help assess whether the models are correctly identifying discontinued products and whether they are prone to misclassification. These metrics are calculated for both the Decision Tree and Random Forest models, allowing for a direct comparison of their performance.

3.6. Feature Importance

One of the advantages of the Random Forest algorithm is its ability to provide feature importance scores, which indicate which variables contribute the most to the model's predictions. We calculate feature importance using the built-in `feature_importances_` attribute of the trained Random Forest model. These scores are then visualized in a bar plot, which ranks the features based on their importance. Features such as `manufacturer`, `price_inr`, and `therapeutic_class` may emerge as the most significant predictors of product discontinuation. Understanding feature importance is valuable because it can guide pharmaceutical companies in identifying which factors influence their product lifecycle decisions the most.

4. Results and Discussion

4.1. Data Overview and EDA Finding

The dataset used in this study consists of 253,973 entries and 15 columns, each representing key attributes of pharmaceutical products available in India. The dataset includes details such as product ID, brand name, manufacturer, price, dosage form, pack size, active ingredients, and the target variable `is_discontinued` (indicating whether the product has been discontinued). Upon initial inspection, it was found that the dataset contained missing values in certain columns, notably `pack_size`, `pack_unit`, and `primary_strength`, which were imputed using appropriate strategies (mean for numerical columns and most frequent for categorical columns). Feature engineering was performed on the `active_ingredients` column, where a new feature, `all_ingredients`, was created by extracting and joining ingredient names for each product. Irrelevant or redundant columns such as `product_id`, `brand_name`, and `primary_strength` were dropped to simplify the dataset and focus on more relevant features for the model.

During the EDA phase, we analyzed the distribution of the target variable `is_discontinued`. The dataset was highly imbalanced, with 96.9% of products labeled as active (`False` for `is_discontinued`) and only 3.1% marked as discontinued (`True`). This imbalance is common in real-world datasets and can affect model performance, as the model may become biased towards predicting the majority class (active products). To further explore the data, we generated several visualizations. A count plot of `is_discontinued` confirmed the imbalance, as shown in figure 1. Additional plots revealed insights into the relationships between manufacturer and discontinuation status, shown in fig 2.

 ISSN 2776-8007

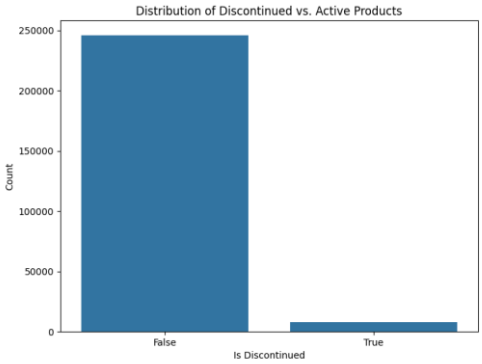

Figure 1. Discontinuation Status Count Plot

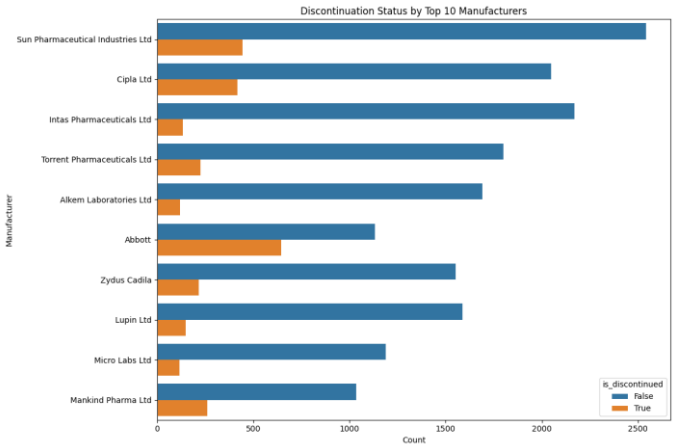

Figure 2. Discontinuation Status by Top 10 Manufacturers

4.2. Results of Model Training and Evaluation

After preprocessing the data and preparing it for modeling, we trained two machine learning models: Decision Tree Classifier and Random Forest Classifier, to predict pharmaceutical product discontinuation. The dataset was split into training and testing sets, with 70% of the data allocated for training and 30% for testing. Stratification was applied to ensure that the distribution of the target variable (`is_discontinued`) was similar in both the training and testing sets, maintaining the balance between active and discontinued products. The Decision Tree Classifier was trained with the default parameters, utilizing Gini impurity as the criterion for splitting nodes. This model, being a single tree, is intuitive and interpretable, but prone to overfitting, especially with deep trees. The Random Forest Classifier, an ensemble method consisting of 100 decision trees, was also trained using the training set. It uses bagging (bootstrap aggregating) to reduce variance and improve generalization by averaging the results of multiple decision trees. The n_jobs=-1 parameter was specified to utilize all available cores for parallel processing, which speeds up model training.

Both models were evaluated using common classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions, while precision and recall provide insights into the performance of the model specifically for the minority class, i.e., discontinued products. The F1-score serves as a balance between precision and recall, particularly useful in imbalanced datasets like this one, where the discontinued products constitute only 3.1% of the total dataset. The Decision Tree Classifier achieved an accuracy of 95.23%, which was relatively high. However, its precision (0.2594), recall (0.2863), and F1-score (0.2721) were low, reflecting the model's difficulty in identifying discontinued products. The model showed a tendency to predict active products more frequently, which is a common issue when dealing with imbalanced datasets. On the other hand, the Random Forest Classifier performed better in terms of accuracy (96.92%), but its recall (0.0982) for discontinued products was quite low, indicating that it was still not very effective at predicting the minority class. Although the precision (0.5271) of the Random Forest model was higher, its low recall and F1-score (0.1656) further demonstrated the challenge of predicting the

 ISSN 2776-8007

discontinued products in the presence of significant class imbalance. Confusion matrix of both models shown in figure 3.

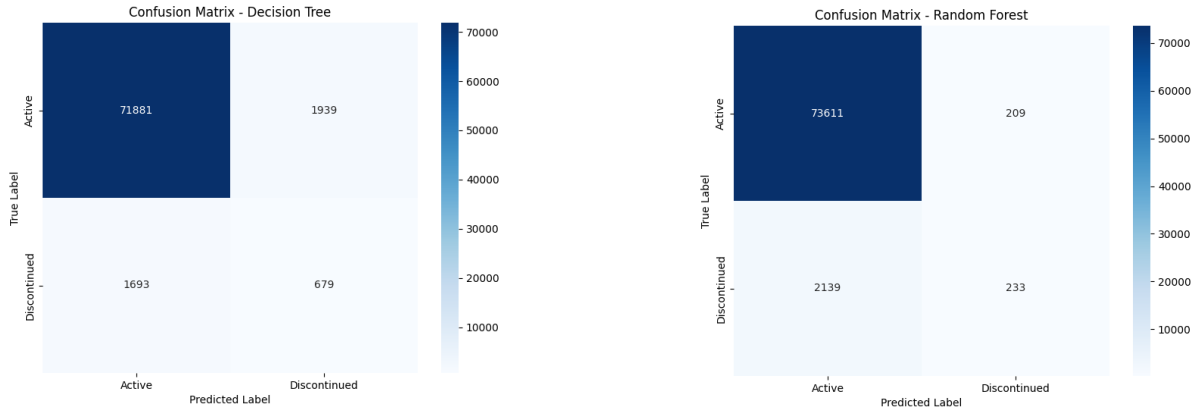

Figure 3. Confusion Matrix of (a) Decision Tree and (b) Random Forest

4.3. Feature Importance Analysis

One of the advantages of the Random Forest algorithm is its ability to compute feature importance, which helps identify which features contribute most to the model's predictions. The feature_importances_ attribute of the Random Forest model provides a ranking of features based on their contribution to the prediction of pharmaceutical product discontinuation, shown in figure 4. From the analysis, the top 5 most important features were price_inr, manufacturer, all_ingredients, primary_ingredient, and pack_size. The feature price_inr had the highest importance score of 0.3300, indicating that the price of the pharmaceutical product is a major factor influencing its likelihood of discontinuation. The second most important feature was manufacturer, with an importance score of 0.3168, suggesting that certain manufacturers are more likely to discontinue products, or perhaps that discontinued products are more likely to come from particular manufacturers.

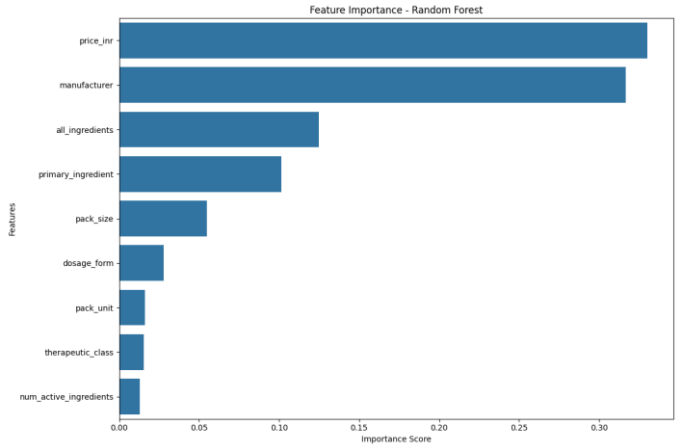

Figure 4. Feature Importance Plot of Random Forest

The feature all_ingredients (which was engineered from the `active_ingredients` column) had an importance score of 0.1249, indicating that the composition of the product plays a significant role in predicting its discontinuation. Other features such as primary_ingredient and pack_size also contributed to the model's predictions, with scores of 0.1014 and 0.0548, respectively. These insights suggest that product attributes such as the key ingredient and packaging size have some predictive value when determining whether a product is likely to be discontinued. The feature importance analysis provides valuable insights into the underlying factors that drive product discontinuation in the pharmaceutical market. It highlights that both the manufacturer and price are the strongest predictors, but ingredient composition and pack size also play notable roles. These findings can help pharmaceutical companies prioritize factors that may reduce the likelihood of discontinuation or focus on improving certain product attributes to extend their market life.

4.4. Discussion

The key findings from this study reveal that certain features play a significant role in predicting whether a pharmaceutical product will be discontinued. Notably, manufacturer and composition emerged as the most influential features. The manufacturer's identity, with an importance score of 0.3168, suggests that products from certain manufacturers are more likely to be discontinued, possibly due to factors such as company strategies, market focus, or product performance. This highlights the importance of brand reputation and market positioning in determining a product's lifecycle. Additionally, the composition of the product, represented by the `all_ingredients` feature, had an importance score of 0.1249, indicating that the specific combination of active ingredients significantly affects a product's survival in the market. Products with more common or competitive ingredients may be less likely to face discontinuation, as they could be more easily replaced by newer alternatives or generic versions.

These findings have important implications for the pharmaceutical industry. Understanding the relationship between manufacturer and product discontinuation can help companies identify which manufacturers are more prone to discontinuing products and why. This insight could lead pharmaceutical companies to re-evaluate their strategies regarding product development, marketing, and inventory management. For example, companies may want to prioritize products from manufacturers with a lower discontinuation rate or invest in better quality control to reduce the likelihood of discontinuation. Additionally, identifying the role of composition can guide companies in optimizing their product formulations to increase their chances of long-term market success. Products with ingredients that are in high demand or essential for particular therapeutic classes may have a better chance of surviving in the competitive market.

The potential for these findings to impact pharmaceutical industry practices is substantial. By incorporating insights from this research, pharmaceutical companies could make more informed decisions regarding product portfolios and lifecycle management. They could apply these findings to improve decision-making processes related to which products to invest in, which to phase out, or which formulations to develop further. Furthermore, these insights could be used to forecast the likelihood of new products being discontinued, allowing companies to adapt their strategies proactively. Overall, the ability to predict discontinuation based on key product features offers valuable opportunities for risk management, cost optimization, and enhancing the overall effectiveness of pharmaceutical product development and marketing strategies.

5. Conclusion

In this research, we developed predictive models to forecast the discontinuation of pharmaceutical products based on key attributes such as manufacturer, composition, price, and packaging. The Decision Tree and Random Forest models were evaluated, with Random Forest achieving the highest accuracy, but both models exhibited challenges in predicting the minority class of discontinued products. Feature importance analysis revealed that manufacturer and composition played the most significant roles in determining whether a product would be discontinued, offering valuable insights into the factors that influence product lifecycles. These findings underscore the importance of both brand reputation and product formulation in predicting discontinuation.

The practical implications of this research are substantial for pharmaceutical companies. By utilizing the predictive model developed in this study, companies can gain valuable insights into their product portfolios and make informed decisions about which products to invest in or phase out. The model can be used to assess the likelihood of product discontinuation, enabling companies to proactively manage their product lifecycles, optimize their portfolios, and allocate resources more efficiently. This approach could help reduce the risks associated with inventory management, marketing, and production costs by identifying products that may be at risk of being discontinued before it happens.

However, there are some limitations to this study. The primary limitation lies in the imbalance of the dataset, with the majority of products being classified as active, which could lead to model bias. Additionally, the models may be constrained by the quality and completeness of the data, such as missing values or imprecise feature definitions. Future research could improve model performance by incorporating additional data sources, such as sales data or market trends, to provide a more comprehensive picture of product discontinuation factors. Furthermore, applying deep learning techniques, such as neural networks or advanced ensemble methods, could potentially enhance prediction

accuracy and handle more complex patterns in the data, ultimately improving the robustness of the model for real-world applications.

6. Declarations

6.1. Author Contributions

Conceptualization: S.H., N.A.; Methodology: S.H.; Software: S.H.; Validation: N.A.; Formal Analysis: S.H.; Investigation: S.H.; Resources: N.A.; Data Curation: S.H.; Writing – Original Draft Preparation: S.H.; Writing – Review and Editing: S.H., N.A.; Visualization: S.H.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. N. Selam, S. Abera, H. Geremew, and E. E. Ali, “Local Pharmaceutical Research and Development Capacity in a Developing Country: A Qualitative Exploration of Perspectives From Key Stakeholders in Ethiopia,” *Journal of Pharmaceutical Policy and Practice*, 2022, doi: 10.1186/s40545-022-00491-3.

[2] C. H. Song and J. Han, “Patent Cliff and Strategic Switch: Exploring Strategic Design Possibilities in the Pharmaceutical Industry,” *Springerplus*, 2016, doi: 10.1186/s40064-016-2323-1.

[3] A. Shabani, S. Rajabi, N. Alipanahi, and A. Ahmadi, “Key Factors to Improve Pharmaceutical Industry’s R&D Productivity: A Case Study of Iranian Pharmaceutical Holding,” *Medical Journal of the Islamic Republic of Iran*, 2022, doi: 10.47176/mjiri.36.117.

[4] A. B. Евстратов, И. А. Езангина, M. V. Shendo, and T. Luneva, “Mergers and Acquisitions on the Pharmaceutical Market: Global Experience and Russian Specifics,” 2019, doi: 10.2991/cssdre-19.2019.12.

[5] Z. Chen, “Risks Analysis and Development Prediction in Pharmaceutical Industry: A Case Study of Johnson and Johnson, Pfizer and Roche,” *Advances in Economics Management and Political Sciences*, 2023, doi: 10.54254/2754-1169/40/20231996.

[6] A. Johnson, I. Fladie, J. Anderson, D. M. Lewis, B. R. Mons, and M. Vassar, “Rates of Discontinuation and Nonpublication of Head and Neck Cancer Randomized Clinical Trials,” *Jama Otolaryngology–head & Neck Surgery*, 2020, doi: 10.1001/jamaoto.2019.3967.

[7] G. Singh, A. Wague, A. Arora, V. Rao, D. Ward, and J. Barry, “Discontinuation and Nonpublication of Clinical Trials in Orthopaedic Oncology,” 2023, doi: 10.21203/rs.3.rs-3707920/v1.

[8] S. Emam, A. X. Du, P. Surmanowicz, S. F. Thomsen, R. Greiner, and R. Gniadecki, “Predicting the Long-term Outcomes of Biologics in Patients With Psoriasis Using Machine Learning,” *British Journal of Dermatology*, 2020, doi: 10.1111/bjd.18741.

[9] S. F. Pratama and A. M. Wahid, "Mining Public Sentiment and Trends in Social Media Discussions on Indonesian Presidential Candidates Using Support Vector Machines," *Journal of Digital Society*, vol. 1, no. 2, Art. no. 2, Jun. 2025, doi: 10.63913/jds.v1i2.8.

[10] Y. Yang, Z. Ye, Y. Su, Q. Zhao, X. Li, and D. Ouyang, "Deep Learning for in Vitro Prediction of Pharmaceutical Formulations," *Acta Pharmaceutica Sinica B*, 2019, doi: 10.1016/j.apsb.2018.09.010.

[11] J. Weygandt *et al.*, "Discontinuation and Nonpublication of Clinical Trials for the Pharmacologic Treatment of Posttraumatic Stress Disorder Among Military Veterans," *Journal of Traumatic Stress*, 2023, doi: 10.1002/jts.22911.

[12] C. Zeng, Y. S. Lee, A. Szatrowski, D. Mero, and B. B. Khomtchouk, "Computational Integration and Meta-Analysis of Abandoned Cardio-(Vascular/Renal/Metabolic) Therapeutics Discontinued During Clinical Trials From 2011 to 2022," *Frontiers in Cardiovascular Medicine*, 2023, doi: 10.3389/fcvm.2023.1033832.

[13] T. Hafner, H. Walkowiak, D. Lee, and F. Aboagye-Nyame, "Defining Pharmaceutical Systems Strengthening: Concepts to Enable Measurement," *Health Policy and Planning*, 2016, doi: 10.1093/heapol/czw153.

[14] C. Pitchayajittipong *et al.*, "An Overview of Pharmaceutical Production in Thai Hospitals," *Hospital Pharmacy*, 2019, doi: 10.1177/0018578719890090.

[15] E. W. Conder *et al.*, "The Pharmaceutical Drying Unit Operation: An Industry Perspective on Advancing the Science and Development Approach for Scale-Up and Technology Transfer," *Organic Process Research & Development*, 2017, doi: 10.1021/acs.oprd.6b00406.

[16] A. Alzahrani and A. Safhi, "The Role of Data Mining Techniques and Tools in Big Data Management in Healthcare Field," *Sustainable Engineering and Innovation Issn 2712-0562*, 2022, doi: 10.37868/sei.v4i1.id128.

[17] A. Pika, M. T. Wynn, S. Budiono, A. H. M. Hofstede, W. M. P. Aalst, and H. A. Reijers, "Privacy-Preserving Process Mining in Healthcare," *International Journal of Environmental Research and Public Health*, 2020, doi: 10.3390/ijerph17051612.

[18] R. R. de Oliveira, S. L. Barbosa Santos, and R. J. Sassi, "Proposed Use of Data Mining Techniques in Healthcare for Knowledge Generation in Patient Care: An Integrative Review," 2023, doi: 10.56238/homeiisevenhealth-054.

[19] C. Diriba, "Developing Risk Level Prediction Model and Clinical Decision Support System for Cardiovascular Diseases in Ethiopia," *International Journal of Clinical Case Reports and Reviews*, 2023, doi: 10.31579/2690-4861/311.

[20] N. M. Zawad, "Application of Data Mining in Healthcare of Bangladesh," *Ijiscs (International Journal of Information System and Computer Science)*, 2023, doi: 10.56327/ijiscs.v7i2.1433.

[21] G. Biau and E. Scornet, "A Random Forest Guided Tour," *Test*, 2016, doi: 10.1007/s11749-016-0481-7.

[22] G. Dudek, "A Comprehensive Study of Random Forest for Short-Term Load Forecasting," *Energies*, 2022, doi: 10.3390/en15207547.

[23] D. Liu *et al.*, "Optimisation and Evaluation of the Random Forest Model in the Efficacy Prediction of Chemoradiotherapy for Advanced Cervical Cancer Based on Radiomics Signature From High-Resolution T2 Weighted images," *Archives of Gynecology and Obstetrics*, 2021, doi: 10.1007/s00404-020-05908-5.

[24] T. Yoshida *et al.*, "Kinetic Estimated Glomerular Filtration Rate as a Predictor of Successful Continuous Renal Replacement Therapy Discontinuation," *Nephrology*, 2019, doi: 10.1111/nep.13396.

[25] G. Loy-García, R. Rodríguez-Aguilar, and J. A. Marmolejo-Saucedo, "An Analytical Intelligence Model to Discontinue Products in a Transnational Company," 2021, doi: 10.1007/978-3-030-68154-8_70.

[26] M. Alghamdi *et al.*, "Developing a Machine Learning Model With Enhanced Performance for Predicting COVID-19 From Patients Presenting to the Emergency Room With Acute Respiratory Symptoms," *Iet Systems Biology*, 2024, doi: 10.1049/syb2.12101.

[27] H. S. Yang *et al.*, "Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning," *Clinical Chemistry*, 2020, doi: 10.1093/clinchem/hvaa200.

[28] M.-S. Chen and S.-H. Chen, "A Data-Driven Assessment of the Metabolic Syndrome Criteria for Adult Health Management in Taiwan," *International Journal of Environmental Research and Public Health*, 2018, doi: 10.3390/ijerph16010092.