University College Cork

2025-IS6611: Applied Research in Business Analytics

IT Artefact Version-3

GENOVIX

A Pharma Intelligence Dashboard Empowering Generic Drug Manufacturers

by

Group – 23

Brihat Kaleru – 124112779
Naresh Kumar Dugginapalli – 124104289
Preeti Chandrakant Jadhav – 124115068
Indhraja Vara Praharika Peta – 124118983
Darun Sundra Selva Kumar – 124103195
Nitish Sugali Mudavath – 124112424

Table of Contents

Introduction

In a regulated industry like pharmaceuticals, data-driven decision-making plays a crucial role in planning market entry, regulatory filing and gaining a competitive advantage (DDReg Pharma, 2024). Historical regulatory data is key in supporting critical processes across the pharmaceutical lifecycle, ranging from the development of new drugs to the traceability of their distribution (FDA, 2005) and to ensure that pharmaceutical products are safe to use and meet the required public health standards (FDA, 2023).

Generic drugs, by offering cost-effective alternatives to brand-name drugs, significantly improve affordability and accessibility, thereby reducing healthcare costs and improving patient access to essential treatment (Miller, 2020). In the US, generic drugs account for every nine out of ten prescription drugs dispensed and in the last five years, generic market revenue has grown to an estimated $50.3 billion, including expected growth of 3.9% in 2025 (IBISWorld, 2025).

Despite this, generic drug manufacturers face challenges in identifying the right opportunities for market entry due to the disjointed nature of FDA approval trends, exclusivity expirations, and therapeutic equivalence gaps data (Brookings, 2017). Drug selection for generic production is a highly manual, time-consuming, and decentralized process. Manufacturers need to search through thousands of disjointed records scattered across multiple FDA databases, the Orange Book, the NDC directory, and shortage lists—none of which are linked with each other. They spend a considerable amount of time aggregating raw data and cleaning and validating data and often find information like therapeutic class or submission filing dates of applications missing. In addition, it is challenging for manufacturers to stay up to date with regularly updating FDA data using the existing manual process.

Key concerns of Stakeholder

The current "as-is" process of finding a potential drug for generic entry involves manufacturers performing all the data-heavy lifting manually, which involves extracting data from FDA sources, consolidating and cleaning data into spreadsheets, followed by performing evaluations drug-by-drug. Due to multiple manual touchpoints, this process is prone to errors and missed opportunities.

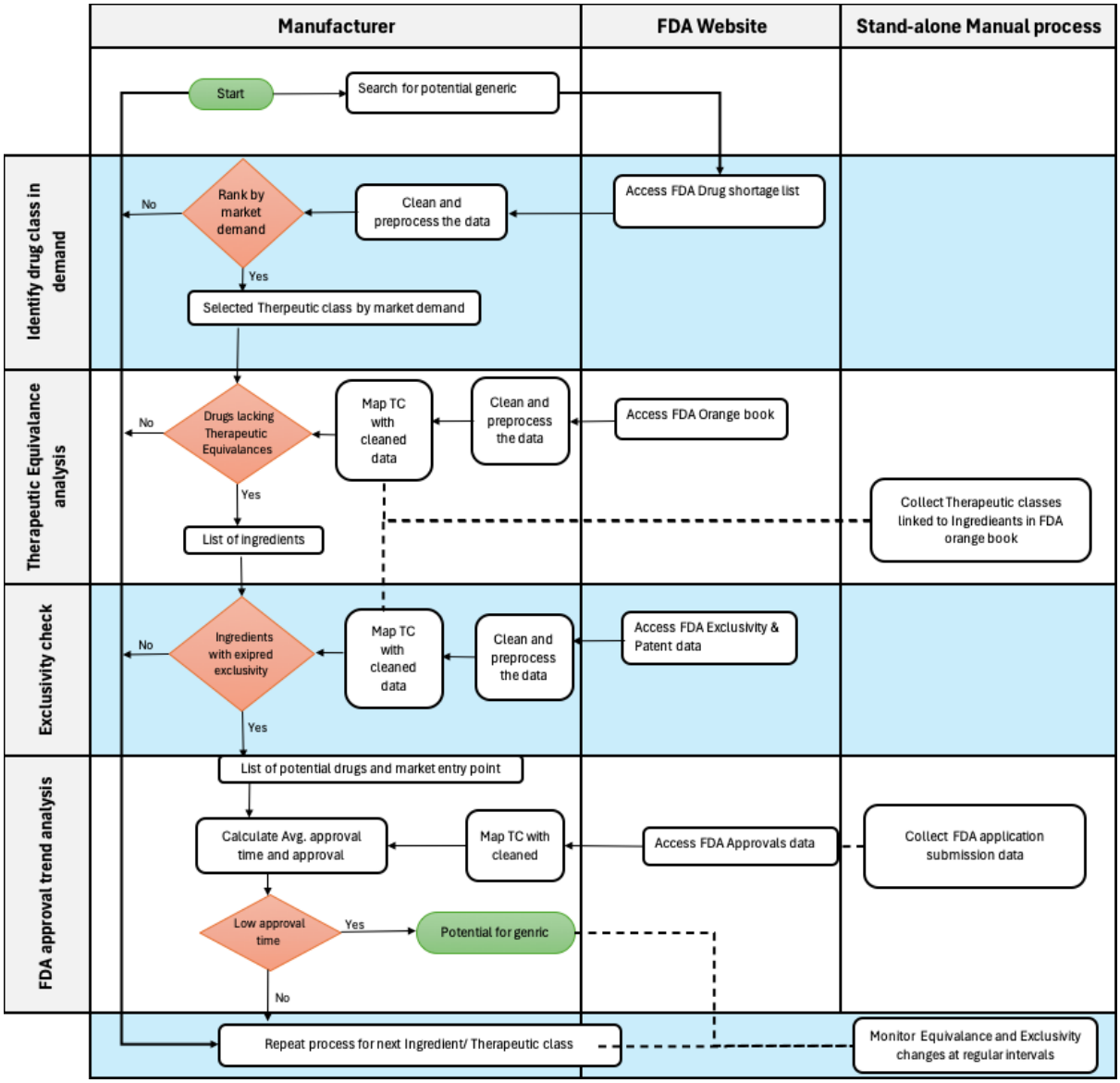

Figure 1 : As-is process swim-lane diagram

In addition to this, the process is not optimal for regularly updating FDA trends, and it also does not support cross-comparative analysis (see figure-1). The lack of FDA submissions data further complicates the process and reduces the level of confidence in the resulting insights, which in turn affects the manufacturer's decision-making ability (DDReg Pharma, 2023).

Process Transformation

GENOVIX is a web-based platform designed to support generic drug manufacturer's decision-making in identify market trends and potential generic drugs. By promoting generics, Genovix contributes to SDG-3 by helping enhance access to affordable essential medicines (Banik, 2019). It replaces inefficient, error-prone manual processes with a unified platform for a seamless analysis of approval trends, market opportunities, competition and identifies regulatory bottlenecks. By consolidating fragmented data sources, Genovix improves the speed

and accuracy of decision-making processes (McCaman Taylor, 2020). The analytical lifecycle and business value of Genovix are depicted in the Data Value Map framework (Nagle & Sammon, 2017) in figure-2.

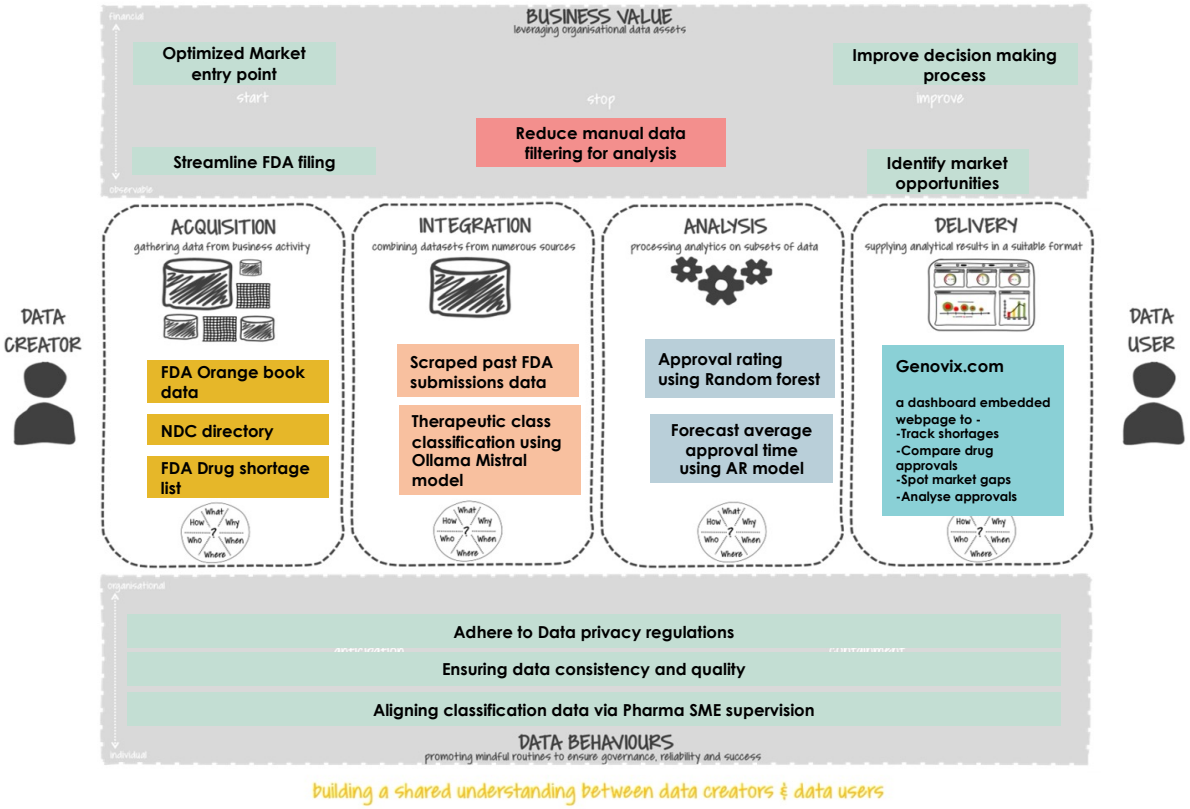

Figure 2: Data Value Map (DVM)

The process begins by acquiring data from the FDA's Orange book, NDC directory and Drug shortage list. This data is then integrated with the FDA application submission data and therapeutic classes, to ensure data consistency and uniformity across different elements of the solution.

Resource/Tool	Used for	Reason for Choice
Ollama Mistral 7B model	Act as a pharma SME to classify drug ingredients into therapeutic classes	Open-source and lightweight AI model, with fast runtime and good accuracy.
Supabase DB	Data storage and transformation	Open-source, it offers a wide range of API connections for seamless connectivity.
Tableau	Data visualization	Easy connectivity to Supabase DB; can be developed and deployed on both macOS and Windows.
Python	Data cleaning, processing, and predictive modelling	Availability of a wide range of libraries like Scikit-learn, pandas, and NumPy.

Table-1 : Resources and Tools

The data extracted is analysed for therapeutic equivalence gaps, drug shortages and to predict the average time using machine learning model and forecasting model, for new applications. Results from the analysis phase are then compiled into key metrics. Insights from these metrics are then delivered to the user using a web-based platform embedded with interactive dashboards to facilitate real-time analysis. Resources and tools used to build the Genovix platform are listed in table-1 along with the key reason for the choice.

Solution Architecture

Genovix is designed with three core principles in mind scalability, flexibility and robust architecture. The web platform is a dashboard embedded website with an interactive, easy to navigate, and user-friendly interface. Instead of offering a specific application based dashboard which results in application dependency on the user end, the solution hosts the dashboard on a website with real-time updates. This approach enables scalability and robustness, allowing access across devices at any time (TryCyfer, 2024).

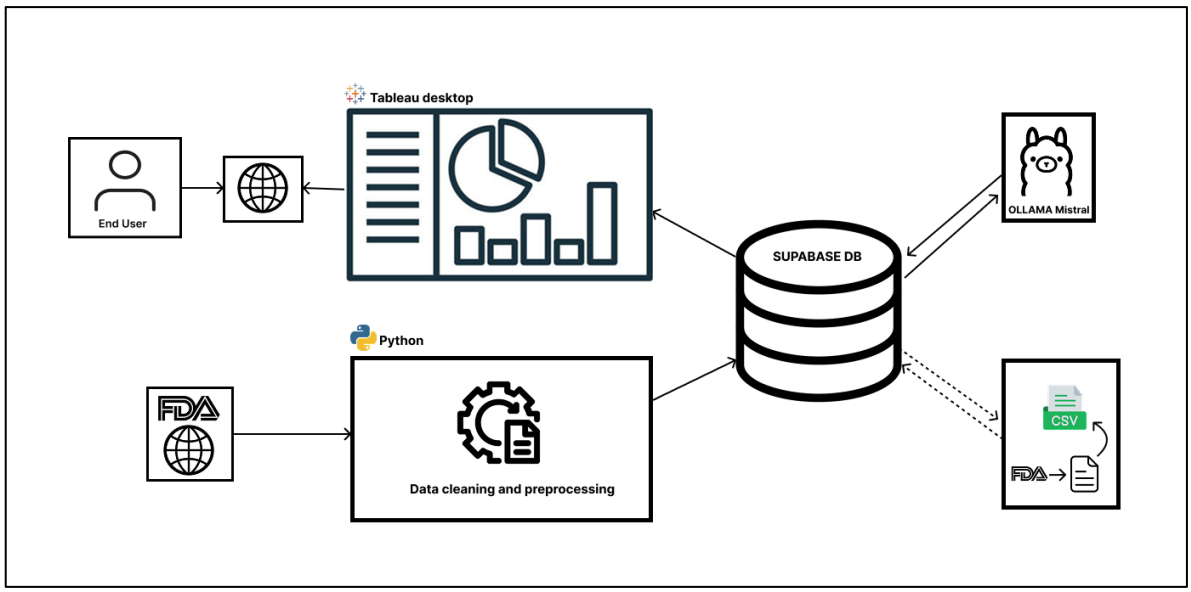

Figure 3 : Technology Architecture (High level view)

Data Acquisition

The Genovix solution leverages publicly available data from FDA's Orange book, NDC directory, and drug shortage list to identify generic opportunities and market entry points. Although the data from these data sources are comprehensive, they lack a crucial field: the therapeutic class of the drug ingredient involved in each drug application(McCaman Taylor,

2020). Additionally, the inconsistent format across the data sources and limited access to FDA application submission data present challenges for creating a complete and structured database. Due to the lack of publicly available datasets for FDA application submission data, Genovix uses manually scraped submission date and submission category data from the FDA approval letters. To account for applications with missing approval letters, the yearly average approval time of the corresponding drug type is used in approval time calculation (Seoane-Vazquez, , 2024). This manually scraped data is integrated into the database using a CSV upload.

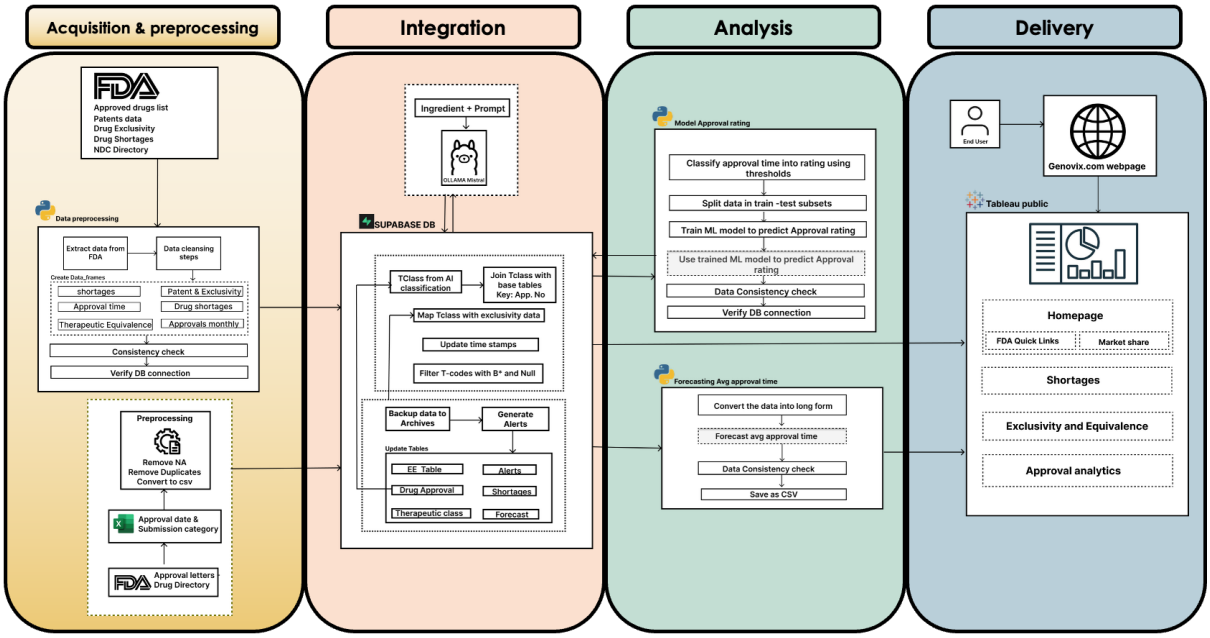

Figure 4 : Technology Architecture (Detailed view)

Integration

To further reduce manufacturers manual effort, we automated the process of extraction, cleaning and data preprocessing using Python. The preprocessing Python script (see Appendix B) is responsible for data extraction from the FDA, removing null values, making the format consistent and preprocessing data into structured sections for loading it into database tables using an API connection.

The database used is an open source PostgreSQL database called ‘Supabase’, it seamlessly connects with all the different modules of the solution using API calls (see figure-3). These API based connections, along with the PostgreSQL based structure, make the database and system architecture both flexible and scalable in nature.Before each database load, a backup step is triggered to archive data from base tables. This not only makes the process robust but also enables the platform to model “Alerts”, which is one of the key components of the solution.

To address the lack of availability of therapeutic classes, the solution uses the OLLAMA MISTRAL 7b model (Parmar, 2024), to classify drug ingredients into their respective therapeutic classes, using a predefined list of therapeutic classes (see appendix). This model was chosen based on its lightweight architecture, strong performance with minimal computational resources and its local deployment capability that ensures data privacy. To ensure consistency with each data refresh, the same AI model is integrated with the database. In each refresh, entries without a therapeutic class are filtered and are classified accordingly using the same predefined list (see figure-4).

Analysis Layer

There are two analytical engines working as part of the Genovix solution (see figure-4). One, the machine learning model used to model "approval rating" and the other is the forecasting algorithm used to predict average approval time for the next five years.

Modelling Approval Rating

The approval Rating is designed to classify each submission—at the levels of Therapeutic Class, Submission Type, Drug Type, and Sub-Type—into one of three categories: High, Medium, or Low. The classification uses the average approval time at the Sub-Class level as a baseline (refer to Appendix).

To achieve this, five machine learning classification models were trained on manually scraped submission data. Each entry in the scraped dataset was first labelled into one of the three categories based on the following logic:

- If the approval time is greater than the baseline plus 30 days, it is classified as low.
- If it is less than the baseline minus 30 days, it is classified as high.
- Any other case is classified as medium.

Once labelled, the data was used to train and validate the models using a cross-validation approach, followed by prediction on unseen data.

The ML models trained and tested as part of the Genovix build were logistic regression, Support Vector Machine (SVM), Decision Tree, Random forest and K-Nearest neighbour (KNN). And the input features used are - Therapeutic class, Submission type, drug type, sub type and Year – and the output is approval rating.

Out of 5 classification models tested, three models - Decision Tree, Random Forest and KNN – performed well with an average accuracy, precision and F1-score of around 80% (see figure -5). To select the ideal model for our use case, these three models are further compared across various key factors to determine the best model with a balanced performance and flexibility.

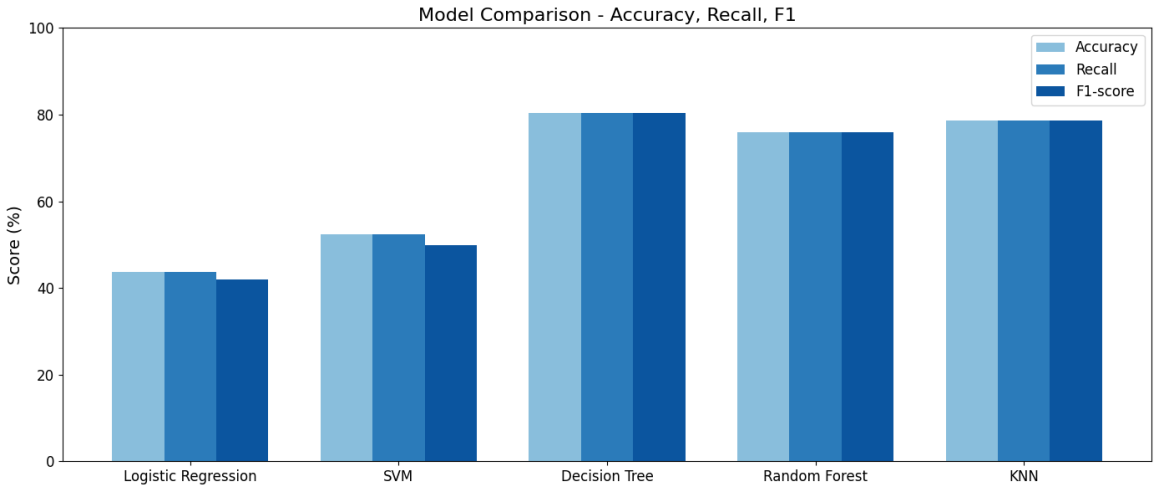

Figure 5 : ML model performance comparison

Though the decision tree has a higher performance compared to KNN and random forest it is prone to overfitting (Al taei, 2024), cannot handle imbalanced data and is inconsistent with results. On the other hand, KNN has a better performance than random forest, but it is not ideal for larger datasets due to its lazy learning approach, making it not suitable for scalable applications (Pramoditha, 2024).

Factor	KNN	Decision Tree	Random Forest
Larger dataset	❌	✅	✅
Accuracy	78%	80%	76%
Fast prediction	❌	✅	✅
Scalability	❌	✅	✅
No Overfitting	✅	❌	✅
Imbalanced classes	❌	❌	✅
Interpretability	❌	✅	❌
Consistent performance	❌	❌	✅

Table-2 : ML model comparison across key factors

This leaves random forest as an ideal choice, from Table-2 it is clear that it has a balanced performance, consistency and is scalable in nature. The feature importance of each input feature is depicted in figure-6 which shows that therapeutic class, year and sub class are the key

determining factors for approval rating. This clearly shows that the model has captured the temporal distribution of data and successfully classified data accordingly.

Finally, the approval Rating for the year 2025 is modelled using a random Forest algorithm across all available feature combinations and stored in the database. This enables quick response times, making the solution independent of model runtime.

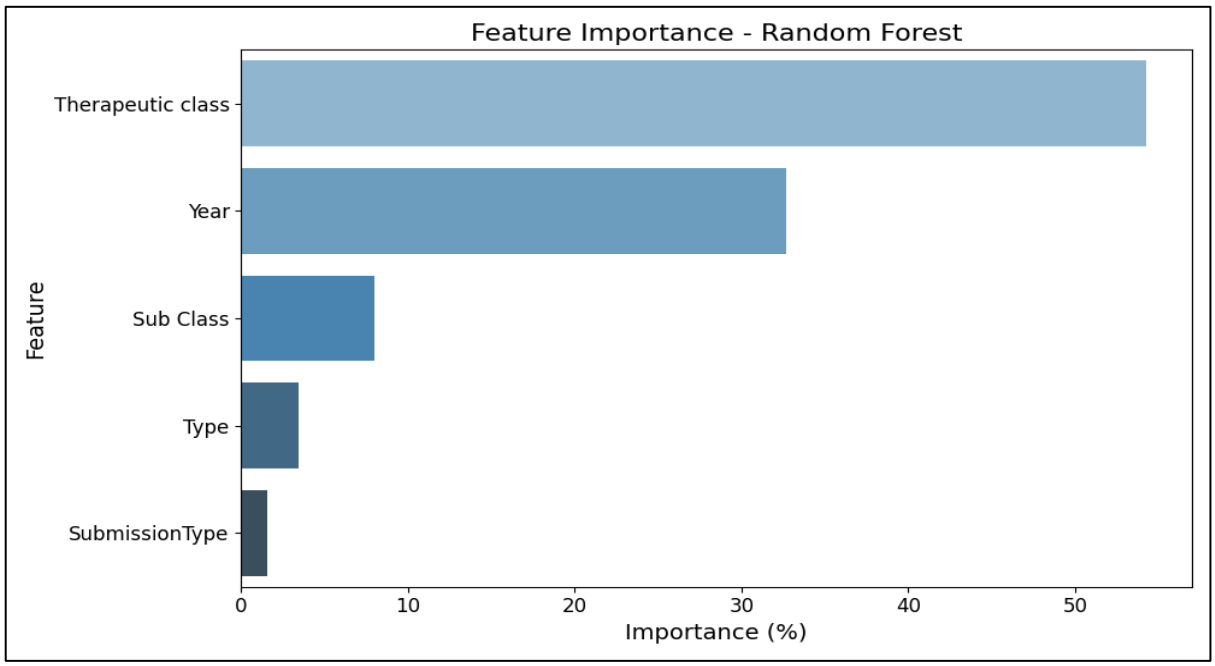

Figure 6: Feature importance Random forest

Forecasting average approval time (days)

The second analytical engine working as part of Genovix is the average approval time forecasting algorithm to predict the average approval time at therapeutic class level. To do this, the scraped data is first grouped together at -therapeutic class, year – level followed by normalizing any extreme outliers. Then the missing values are filled using the average approval time, followed by a stationarity check using the Augmented Dickey-Fuller (ADF) test. Therapeutic classes that are found non-stationary are differenced to make the series stationary, followed by using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood to determine the best one out of the 4 models tested – Auto Regression (AR), ARIMA and SARIMA. Table-3 shows the average of performance metrics across all the therapeutic classes.

Metric	AR(2)	ARIMA	SARIMA
Average AIC	-861.72	-391.94	-106.98
Average BIC	-857.96	-387.81	-101.99
Average LogLik	-111.08	-122.4	-127.5

Table 3 : Forecasting model evaluation metrics

Out of the three models, the AR(2) model consistently showed the lowest AIC/BIC and the highest log-likelihood, making it the best overall model for forecasting approval days (APXML, 2024). Data from these two analytical engines is sent to the database at a detailed level to facilitate visualizations.

Delivery

To deliver insights efficiently, an interactive webpage comprising four key sections - home, shortages, equivalence & exclusivity, and approval analytics - with embedded dashboards has been selected as the delivery format. The webpage "Genovix.com" was built using HTML5 and all the interactive features were modelled using CSS script. And to configure the embedded dashboard, although Power BI offers strong data visualization capabilities and supports PostgreSQL connections, Tableau was chosen for this project because it provides more seamless integration with Supabase using API connections, whereas Power BI requires additional setup to achieve the same level of connectivity. This seamless connection, extraction, and real-time visualization capability without requiring additional configuration ensured the scalability and robustness of the overall solution.

Impact on Stakeholders

Figure-7 shows how Genovix can help manufacturers find and time market opportunities for generic entry. With the use of Genovix, all the data heavy lifting is automated, and combining all the fragmented data sources under a single platform allows comparison of different aspects of the drug data simultaneously, improving flexibility in analysis.

As the dashboard keeps updating in the background at regular intervals, checking for changes in approval or exclusivity status manually is no longer needed. The alert feature of the dashboard makes it easy for the manufacturer to keep track of all the important changes in therapeutic codes, patent and exclusivity expiry dates. Overall, Genovix makes the whole

process faster, smarter, and automated, helping users concentrate on the decision making rather than data cleaning and monitoring.

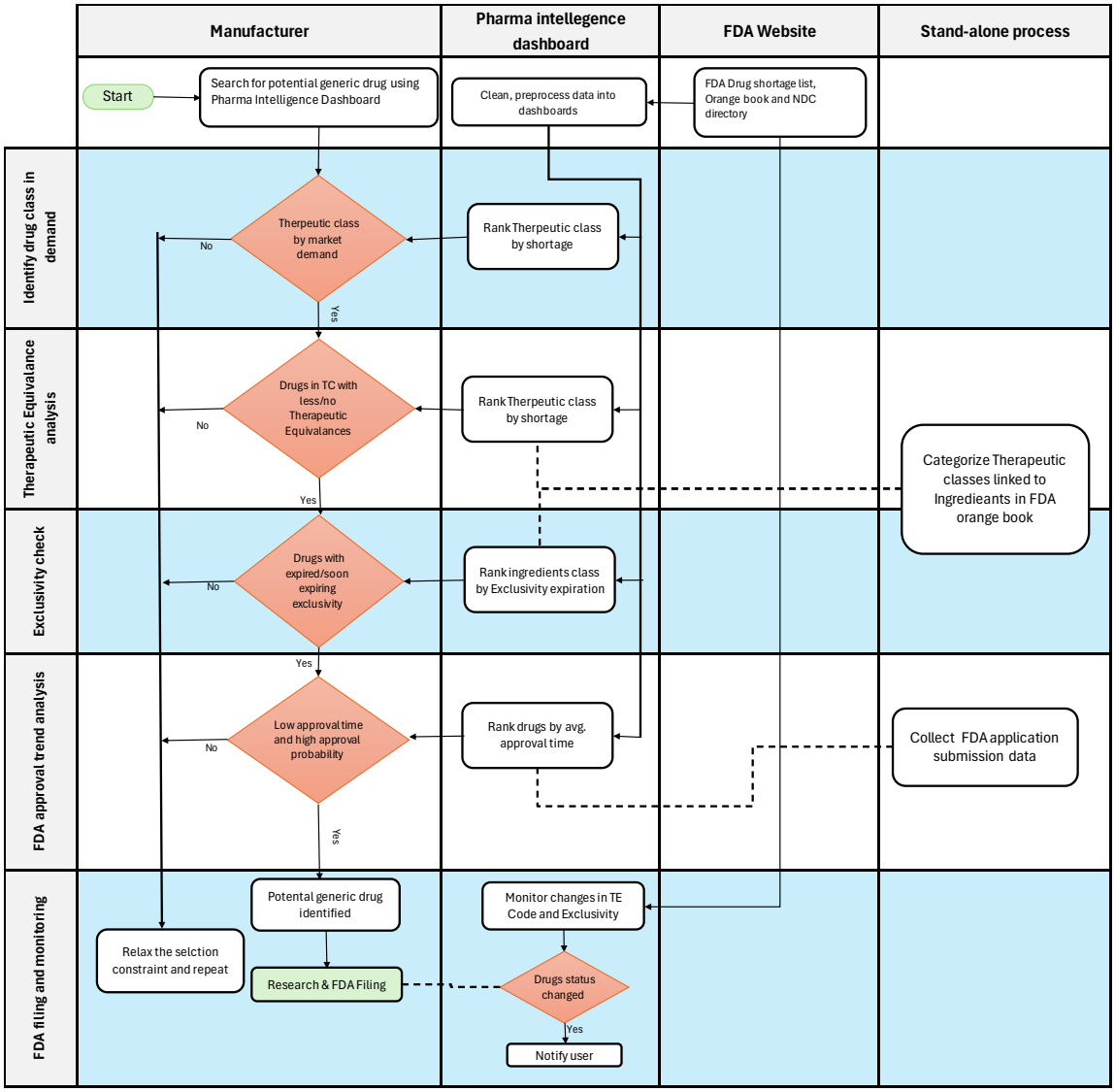

Figure 7 : To-be process swimlane diagram

Business Value

Genovix significantly enhances business value by transforming how generic drug manufacturers identify and capitalise on market opportunities. By integrating fragmented FDA data sources into a unified, automated platform, it eliminates time-consuming manual processes and provides real-time insights into approval trends, therapeutic equivalence gaps, and exclusivity timelines. This allows stakeholders to make data-driven decisions with greater confidence and speed. The platform reduces regulatory submission risks by streamlining data monitoring, improving accuracy, and supporting strategic timing of FDA applications. It also

tracks the therapeutic classes with high approval time, thereby helping note FDA bottlenecks and possible delays. This not only helps manufacturers plan better for FDA filing but also helps them optimize resource allocation and time the market entry accurately.

Through a self-service dashboard with alerts on regular updates in FDA data, Genovix enhances transparency and enables manufacturers to make proactive decisions (DrugPatentWatch, 2025). Its flexible architecture, powered by AI-based classification and approval trend modelling, drives operational efficiency by promoting innovation in the generic drug space.

Genovix ensures scalability while providing broader access to insights for a measurable business and better public health outcomes. Finally, Genovix is not just a data visualization and approval time prediction platform—it is a scalable and flexible one-stop solution for all FDA data analysis needs, capable of performing comparative analysis and backed by robust analytics to support informed decision-making for manufacturers.

This marks a significant leap in process transformation—from a manual, time-consuming, and error-prone data filtering approach to an automated system that delivers insights instantly with minimal manual intervention.

Core Functionality of Genovix

The platform is divided into 4 sections (see figure-8) – home, shortages, equivalence & exclusivity, approval analytics – for the streamlined navigation and targeted analysis of pharmaceutical data.

Figure 8 : Genovix platform navigation

- **Home** provides an overview dashboard summarizing key metrics across all sections.
- **Shortages** tracks current and historical drug shortages, enabling users to identify drug classes with higher demand.
- **Equivalence & Exclusivity** highlights therapeutic equivalent options and monitors upcoming patent or exclusivity expirations to identify potential generic opportunities.
- **Approval Analytics** offers insights into drug approval timelines, application trends, and regulatory patterns by drug class and submission type.

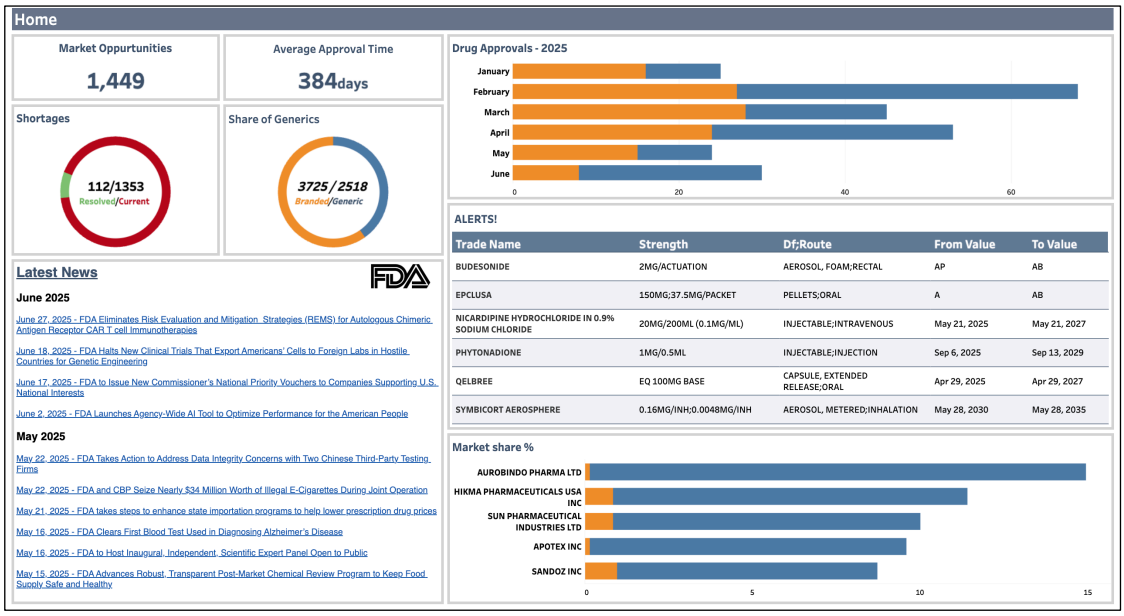

Figure 9 : Genovix homepage

The homepage of Genovix provides an overview of key metrics from all other sections of the platform. It includes quick links to the FDA's recent announcements, drug approvals from the last six months, and information about major companies involved in FDA-approved drugs. One important feature on this page is the "Alerts" table (see figure 9), which shows updates in therapeutic equivalence (TE) codes, patent expiry dates, and exclusivity expiry dates. These alerts are generated by comparing the most recent data with previously stored data each time the system is updated.

Figure 10 : Genovix shortage page

The Shortages section of Genovix translates directly to market demand (see Figure 10), displaying the number of ongoing and resolved shortages along with the yearly distribution of shortages at the therapeutic class level. By hovering over each therapeutic class, users can view the primary reason behind the shortage. For example, in Figure 10, we observe that for cardiovascular drugs, the major reason for shortages is a rise in demand, making it an optimal class to explore. The table on this page depicts a comprehensive list of shortages in each therapeutic class, including details on dosage forms and the companies responsible.

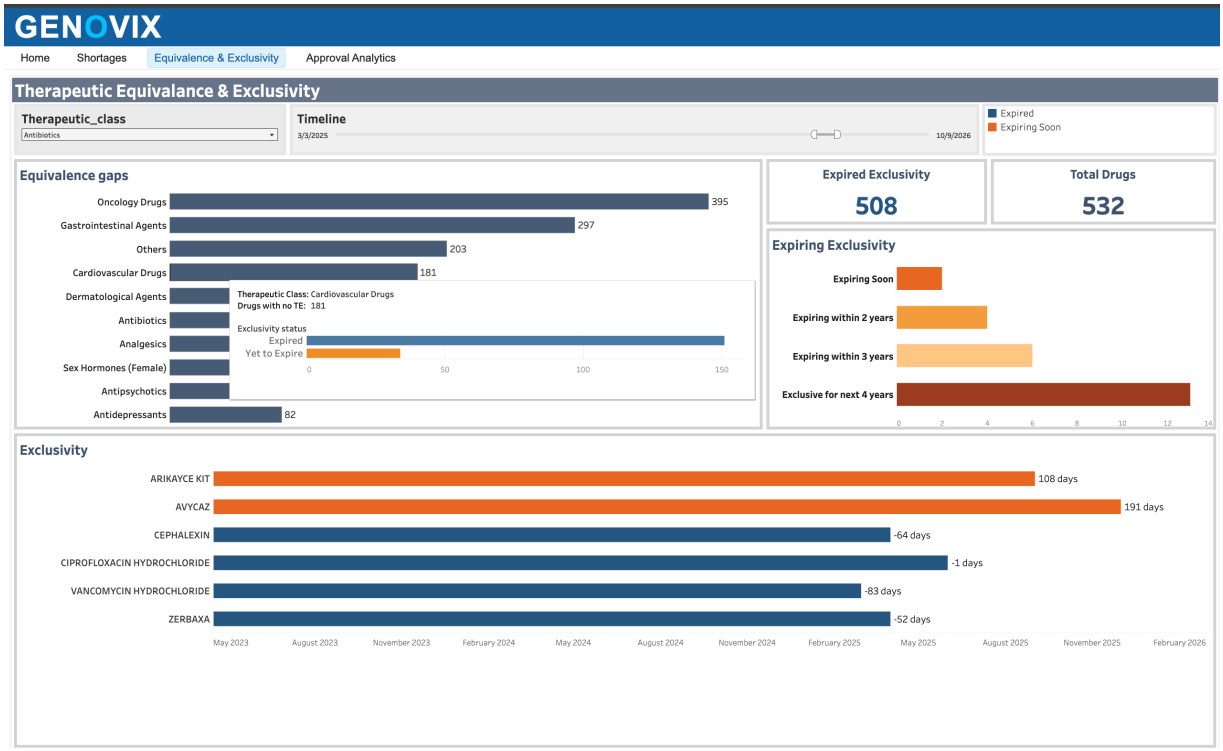

Figure 11 : Genovix Equivalence & Exclusivity page

The Equivalence and Exclusivity page helps users understand available market alternatives and FDA exclusivity restrictions in the selected therapeutic class, aiding in the assessment of market competition and timing of the FDA filing process. This page includes key KPIs such as the total number of approved drugs in the selected therapeutic class, the number of drugs with expired exclusivity, and a bar graph displaying when the remaining drugs are scheduled to lose exclusivity(see figure 11).

The Equivalence bar graph shows the total number of drugs in the selected therapeutic class that lack therapeutic equivalents, indicating the level of market competition. On hovering over the graph, users can view how many of these drugs without therapeutic equivalence have had their exclusivity expired, this is crucial for identifying potential generic drug opportunities. A

timeline slider at the top of the page allows users to filter drugs based on exclusivity expiration dates allowing for a focused drug level analysis.

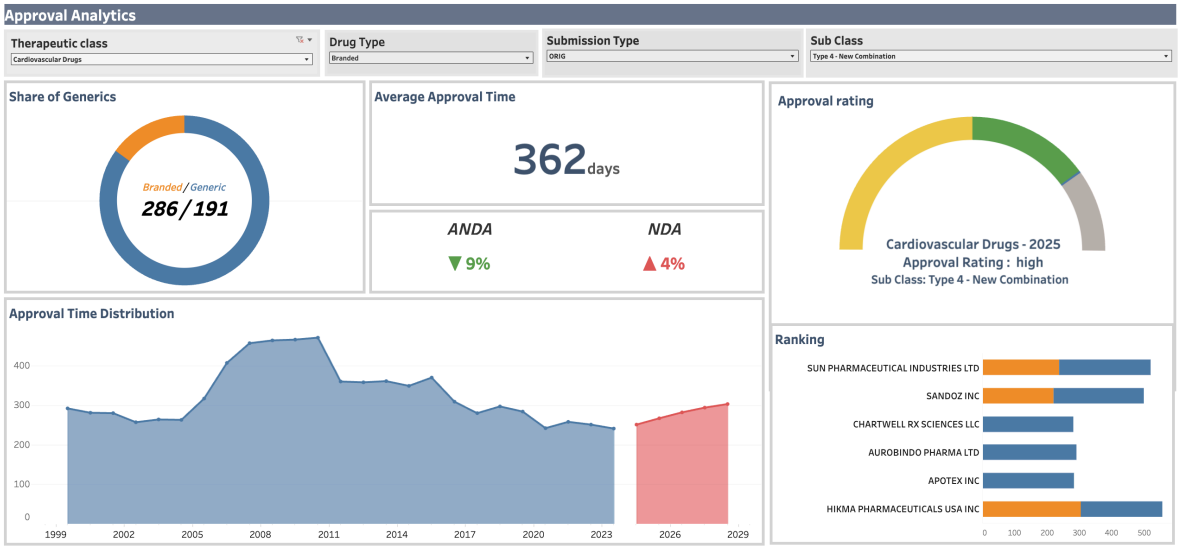

Figure 12 : Genovix Approval analytics page

The approval analytics page(see figure-12) of the Genovix is targeted to help the user make an informed decision about the market entry of the generic drug. This page depicts the average FDA approval time for the selected class along with the year-on-year change in approval time for both generic and brand named drugs, along with the total share of share of generics vs branded drugs in that class(see figure 13).

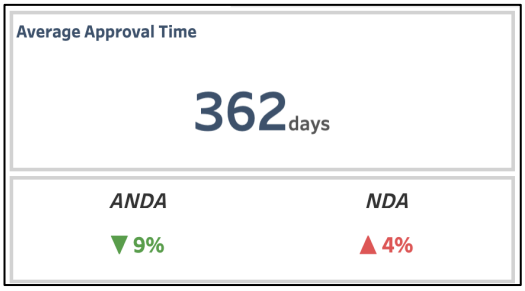

Figure 13 : Genovix Approval time KPI

This page not only helps the user to get an estimated market entry date but also shows the confidence in the data displayed using the results from ML model in the Approval rating tile.

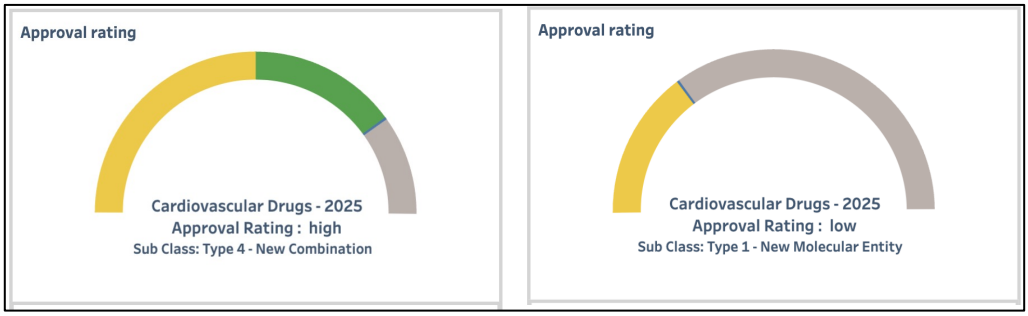

Figure 14 : Genovix Approval rating

The Approval Rating tile is a gauge chart designed to display one of three outcomes - high, medium, or low - based on the selected therapeutic class, submission type, subclass, and drug type. For example, from Figure 14, we can observe that within the same drug class - Cardiovascular Drugs - the approval rating is high for the subclass "Type 4 - New Combination," whereas it is low for the subclass "Type 1 - New Molecule Entry." This implies that for cardiovascular drugs, the approval time for a new molecule entry is higher than the subclass average, meaning it would take longer to receive FDA approval. In contrast, for new combinations, the rating is high, indicating that the FDA approval time is within the subclass average and therefore more likely to receive approval within that timeframe.

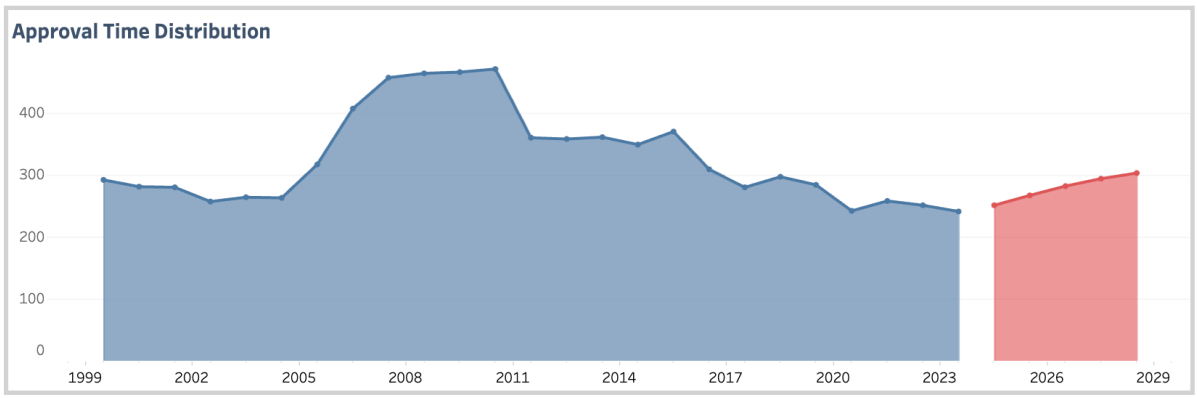

Figure 15 : Genovix Average approval time distribution

The Approval Time Distribution graph (see figure-15) on this page represents the distribution of average approval times for the selected class over the past 20 years and the next 5 years. Data prior to the year 2000 has been filtered out, as 2000 marks the introduction of electronic submissions by the FDA (Applied Clinical Trials, 2001). Heavy outliers in the data were addressed by capping them at a threshold value, followed by forecasting using an AR(2) model.

Conclusion

Genovix offers a data-driven, scalable solution to transform how generic drug manufacturers identify untapped market opportunities. By integrating FDA's fragmented datasets, automating therapeutic class classification, and tracking exclusivity expirations and therapeutic equivalence gaps, the platform reduces the time and effort required to evaluate the viability of generic entry. The use of lightweight AI models like Ollama Mistral, paired with an efficient Supabase backend database and Tableau for visualization, ensures that insights are not only accurate but also easily accessible to stakeholders.

Looking ahead, to expand Genovix's capabilities we can start by integrating insights from real pharmaceutical subject matter experts (SMEs) to improve the accuracy of therapeutic class assignments and validate model assumptions & features. Scale up to a larger dataset by incorporating more historical FDA approval and submission data. To achieve seamless automation, we can adopt background job scheduling that continuously monitors FDA datasets for new approvals, exclusivity expirations, and shortage reports, ensuring that insights remain current without manual intervention. These enhancements will make Genovix a robust and adaptive platform, capable of serving a wide range of generic manufacturers and regulatory consultants across the regulatory industry.

The successful development of Genovix became possible by close collaboration across our multidisciplinary team. Each member, by utilizing their unique expertise in various areas of the build (see appendix). By distributing tasks such as data scraping, data pipeline development, model evaluation, and dashboard design, we ensured parallel progress. Regular team meetings combined with mentors feedback helped validate key assumptions. This collective effort allowed us to create a robust, scalable, and flexible solution.

References

Al taei, R. (2024, October 13). *Understanding overfitting in decision trees*. Medium. https://raghda-altaei.medium.com/understanding-overfitting-in-decision-trees-df31926a31b8

Applied Clinical Trials. (2001, October 1). *The evolution of electronic submissions*. *Applied Clinical Trials*. https://www.appliedclinicaltrialsonline.com/view/evolution-electronic-submissions

APXML. (2024). *Information criteria: AIC & BIC*. In *Time series analysis & forecasting – Chapter 6: Model evaluation & selection*. Retrieved July 17, 2025, from https://apxml.com/courses/time-series-analysis-forecasting/chapter-6-model-evaluation-selection/information-criteria-aic-bic

Banik, D. (2019, September 12). *Making medicines available and affordable*. Centre for Development and the Environment, University of Oslo. https://www.globe.uio.no/english/sdg/blog/dan-banik/making-medicines-available-and-affordable.html

BISWorld. (2025). *Generic pharmaceutical manufacturing in the US - Market size, industry analysis, trends, and forecasts*. https://www.ibisworld.com/united-states/industry/generic-pharmaceutical-manufacturing/488/

Brookings. (2017). Ten challenges in the prescription drug market—and ten solutions. Retrieved from https://www.brookings.edu/articles/ten-challenges-in-the-prescription-drug-market-and-ten-solutions/

DDReg Pharma. (2023). Challenges in Complex Generic Drug Development. Retrieved from https://resource.ddregpharma.com/blogs/challenges-in-complex-generic-drug-development/

DDReg Pharma. (2024, October 14). *The role of data-driven insights in enhancing regulatory strategies*. https://resource.ddregpharma.com/case-studies/

DrugPatentWatch. (2025, May 18). *What happens when a drug patent expires? Understanding drug patent life*. DrugPatentWatch. https://www.drugpatentwatch.com/blog/what-happens-when-a-drug-patent-expires/

McCaman Taylor, L. G. (2020, August 28). *Comments on Approved Drug Products with Therapeutic Equivalence Evaluations (the "Orange Book")* [Comment letter]. National Health Law Program. https://healthlaw.org/wp-content/uploads/2020/09/Orange-Book-Comments_formatted_forsubmission.pdf

Miller, S. (2020). *Generic drugs: A treatment for high-cost health care*. Missouri Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC7023936/

Nagle, T., & Sammon, D. (2017). The Data Value Map: A framework for developing shared understanding on data initiatives. In *ECIS 2017: Proceedings of the 25th European Conference on Information Systems* (pp. 1439–1452). Guimarães, Portugal. https://aisel.aisnet.org/ecis2017_rp/93

Parmar, S. (2024, January 17). *Mistral 7B LLM: Run locally with Ollama*. Medium. https://medium.com/@parmarshyamsinh/mistral-7b-llm-run-locally-with-ollama-7d67e21ad57a

Pramoditha, R. (2024, November 13). *Approximate Nearest Neighbors (ANN) vs K-Nearest Neighbors (KNN) for large datasets: Sacrificing some quality in favour of speed*. Medium. https://rukshanpramoditha.medium.com/approximate-nearest-neighbors-ann-vs-k-nearest-neighbors-knn-for-large-datasets-84f4b5361f97

Seoane-Vazquez, E., Rodriguez-Monguio, R., & Powers, J. H., III. (2024). Analysis of US Food and Drug Administration new drug and biologic approvals, regulatory pathways, and review times, 1980–2022. *Nature Reviews Drug Discovery*. https://doi.org/10.1038/s41573-023-00768-w

TryCyfer. (2024, August 12). *What is a dashboard on a website?* https://trycyfer.com/what-is-a-dashboard-on-a-website/

U.S. Food and Drug Administration. (2022). *Drug approvals and databases*. U.S. Department of Health and Human Services. https://www.fda.gov/drugs/development-approval-process-drugs/drug-approvals-and-databases

U.S. Food and Drug Administration. (2023). *What we do*. U.S. Department of Health and Human Services. https://www.fda.gov/about-fda/what-we-do

U.S. Food and Drug Administration. (2005, March). *Good pharmacovigilance practices and pharmacoepidemiologic assessment: Guidance for industry*. U.S. Department of Health and Human Services. https://www.fda.gov/media/71546/download

United Nations. (n.d.). *Goal 3: Ensure healthy lives and promote well-being for all at all ages*. United Nations. https://www.un.org/sustainabledevelopment/health/

Appendix

A. Data Sources

The following datasets were used in this analysis:

File Name	Source
Products.txt	FDA Orange Book products view
drug_shortages.csv	FDA Drug Shortages
Exclusivity.txt	FDA Orange Book exclusivity view
Patents.txt	FDA Orange Book patents view
NDC directory.xls	FDA NDC directory

All these data files are stored in the project repository under the /data/ folder, available at: https://drive.google.com/drive/folders/1TXjsfdWZ5sviUjveQqfXc-aTrTVm5uOq?usp=sharing

B. Tools used

The following tools and libraries were used throughout the project:

- PYCHARM - Python 3.10
- OLLAMA MISTRAL AI model version 0.3
- SUPBASE Database
- Drugs@Fda API

C. Team roles and key contribution.

The table below outlines the key contributions and corresponding team roles.

Phase	Key Contributors	Key Activities
Problem Scoping	Entire Team	Defined the business problem and aligned it with SDG-3
Identifying Datasets	Naresh, Preeti, Brihat	Identified key data elements and mapped sources for analysis
Data Acquisition	Entire Team	Collected Orange Book, NDC, shortage list, and scraped FDA submissions
Integration & ETL	Naresh, Preeti	Developed Python and SQL ETL scripts; loaded cleaned data into Supabase
AI for Classification	Naresh	Used Ollama model as a pharma SME to classify drug ingredients
Analytical Modelling	Naresh, Darun, Nitish, Brihat	Trained ML models and forecasted approval timelines

Dashboard build	Naresh, Preeti, Brihat, Nitish	Designed and tested interactive dashboards
Webpage design	Preeti, Indhraja, Darun	Built interactive webpage and embedded dashboards
Review & Testing	Entire Team	Tested features, validated results, and refined the final solution
Presentations	Indhraja, Naresh, Preeti	Prepared slides and script for in-class presentation and demo
Report Writing	All Members	Wrote, edited, and formatted the report collaboratively
Proofreading & Checks	All Members + AI Tools	Performed grammar and spell checks using human and AI assistance

D. Python code

The python code used to classify ingredients into Therapeutic classes, model Approval rating and data extraction and preprocessing are stored under the /script/ folder, available at: https://drive.google.com/drive/folders/1vvSiMwosNtjY0zIw0ziE89u7rJilkmQK?usp=sharing

E. Prototype

An interactive Genovix webpage with embedded, Tableau dashboards connecting to tableau public is available at: https://drive.google.com/file/d/11BwOllnnI25x7PJBsJfsPidCBeEJgidu/view?usp=sharing