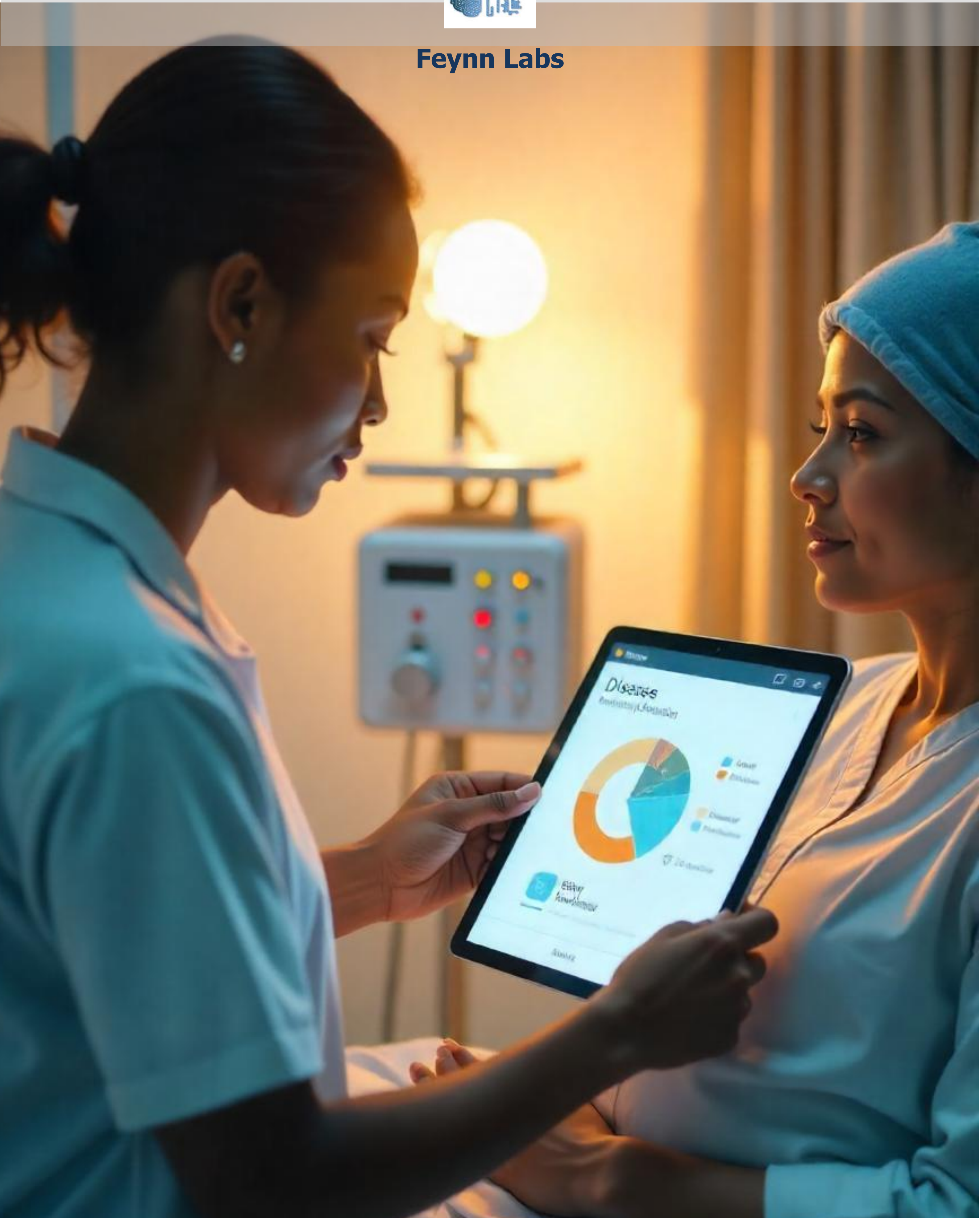


# Feynn Labs

# Disease Prediction System Using Machine Learning

## “AI Diagnosis”

Year 2024

**Authored By**

Nirmal Koshy
Harsh Singh Rana
Anjali
Asna Jamsheed

## 1. Introduction

This report presents the design and development of a machine learning-based disease prediction system. The objective of the system is to assist in predicting diseases based on the symptoms input by users, offering a tool that can be used by both healthcare professionals and individuals to improve diagnostic accuracy and efficiency. The system uses multiple machine learning models, each of which analyses the symptoms and provides a probable diagnosis.

The system was built in a structured manner, beginning with preparing the data, which involves tasks like cleaning and organizing information for better processing. Following this, various machine learning models were trained, tested, and evaluated to ensure that they were producing accurate results. The models used in this system include Support Vector Machines, Gaussian Naive Bayes, and Random Forest Classifiers.

The final prediction process involves combining the predictions of the individual models, and this ensemble method enhances the accuracy and reliability of the predictions. Additionally, a symptom encoding mechanism was created, allowing the system to convert symptom descriptions into a format suitable for the machine learning models to process.

The system is designed to be user-friendly, accessible through an app or web platform, and can be used in a variety of settings such as hospitals, telemedicine platforms, and by individual patients. By leveraging machine learning, this system provides real-time, accurate predictions that could potentially lead to earlier diagnoses and improved patient outcomes

## 2. Data Preprocessing

### 2.1 Loading Data

- **Dataset**: Loaded from a CSV file named "Training.csv".

### 2.2 Handling Missing Values

- **Action**: Dropped any columns with missing values to ensure data integrity.

### 2.3 Label Encoding

- **Target Variable**: The "prognosis" column, containing disease labels, was converted to numerical values using LabelEncoder.

### 2.4 Splitting Data

- **Method**: Used train_test_split to divide the dataset into:
  - **Training Set**: 80% of the data.
  - **Testing Set**: 20% of the data.

### 2.5 Data Balance Check

- A **bar plot** was generated to visualize the distribution of diseases, helping identify whether the data was balanced or imbalanced.

## 3. Model Training and Evaluation

### 3.1 Models Used

Three machine learning models were implemented:

- **Support Vector Machine (SVM)**
- **Gaussian Naive Bayes**
- **Random Forest Classifier**

### 3.2 Cross-Validation

- **10-fold cross-validation** was performed to evaluate each model.
- **Accuracy Scores**: Computed and printed for comparison.

### 3.3 Individual Model Training and Testing

- **Training**: Each model was trained on the training set.
- **Testing**: Models were evaluated on the test set.
- **Accuracy Scores**: Printed for both the training and testing phases.
- **Confusion Matrices**: Generated to visually inspect model performance.

## 4. Combined Model

### 4.1 Final Model Training

- **Models Trained**: The final versions of the SVM, Gaussian Naive Bayes, and Random Forest models were trained on the **entire dataset**.

### 4.2 Prediction on Test Data

- **Prediction**: The trained models made predictions on the test dataset.
- **Final Prediction**: Determined by taking the **mode** of individual model predictions.
- **Evaluation**:
    1. **Accuracy**: The final accuracy of the combined model on the test dataset was calculated.
    2. **Confusion Matrix**: Generated to further evaluate performance.

## 5. Symptom Index Dictionary

### 5.1 Symptom Encoding

- A **symptom index dictionary** was created to map symptom names to numerical indices, making it possible to encode symptoms for model predictions.

## 6. Prediction Function

### 6.1 Input Handling

- The predictDisease function:
    1. Accepts a **string of symptoms** separated by commas.
    2. **Encodes symptoms** and creates an input data array.
    3. Generates predictions from each of the three models.

### 6.2 Final Prediction

- **Final Prediction**: The mode of individual model predictions is taken to determine the final result.
- **Output**: A dictionary is returned containing:
    1. Predictions from each model.
    2. The final combined prediction.

## 7. Testing the Prediction Function

### 7.1 Sample Input Testing

- The function was tested with a **sample input** to verify correctness.
- **Predictions**: Printed to confirm the accuracy of the function's output.

## 8. Problem Statement

The objective of this analysis is to perform market segmentation based on customer demographics, behaviours, and purchasing habits. By dividing the customer base into distinct segments, the company can:

- Tailor marketing strategies, product offerings, and pricing models to the specific needs of each group.
- Enable more effective resource allocation.
- Improve overall customer satisfaction.

## 9. Approach

### 9.1 Dataset

The dataset consists of customer data with key variables, including:

- **Age**
- **Income**
- **Purchasing History**
- **Product Preferences**
- **Geographical Information**

### 9.2 Objective

The goal is to identify meaningful groups that share common characteristics within the data. The segmentation will provide insights into customer behaviours and preferences, enabling:

- Targeted marketing campaigns.

- Tailored product offerings.
- Optimized resource allocation for the company.

### 9.3 Data Preprocessing

- **Handling Missing Data**: Any missing or erroneous data points will be handled through imputation or removal.

## 10. Data Collection

### 10.1 Sources

The data used for this analysis was gathered from multiple sources related to customer demographics and purchasing behaviours:

- **Surveys**: Provided insights into customer preferences.
- **Transactional Records**: Gave a clear picture of purchasing habits and frequency.
- **Customer Profiles**: Included demographic details like age, income, and location.

### 10.2 Variables

Key variables used in the analysis:

- **Age**
- **Income**
- **Geographical Location**
- **Purchasing History**
- **Product Preferences**

### 10.3 Purpose

This combination of data types enables a comprehensive understanding of the customer base, ensuring the segmentation results are accurate and actionable.

## 11. Behavioural Segmentation

### 11.1 Focus

Behavioural segmentation divides the market based on customer behaviours, such as:

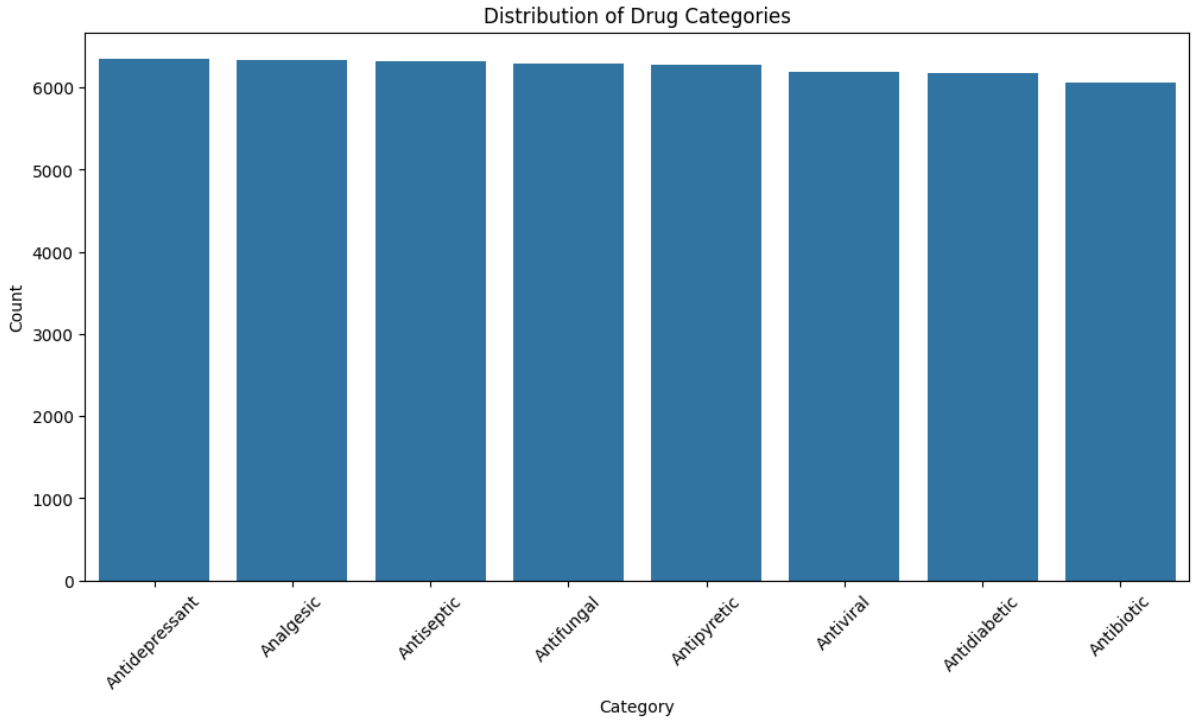
- Purchase frequency.
- Brand loyalty.
- Specific customer needs.

### 11.2 Key Aspects

By studying how customers interact with products, their buying patterns, and their response to marketing efforts, companies can:

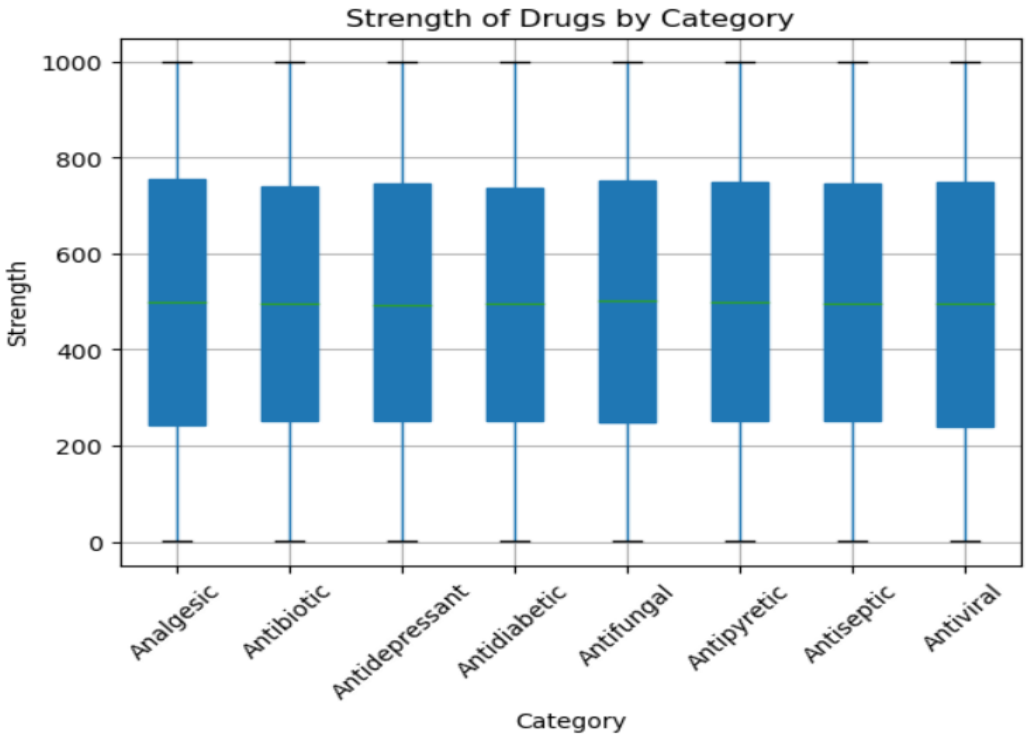
- Focus on customers' readiness to purchase.
- Understand usage rates.
- Uncover motivations behind purchasing decisions.

## 12. Data Visualization

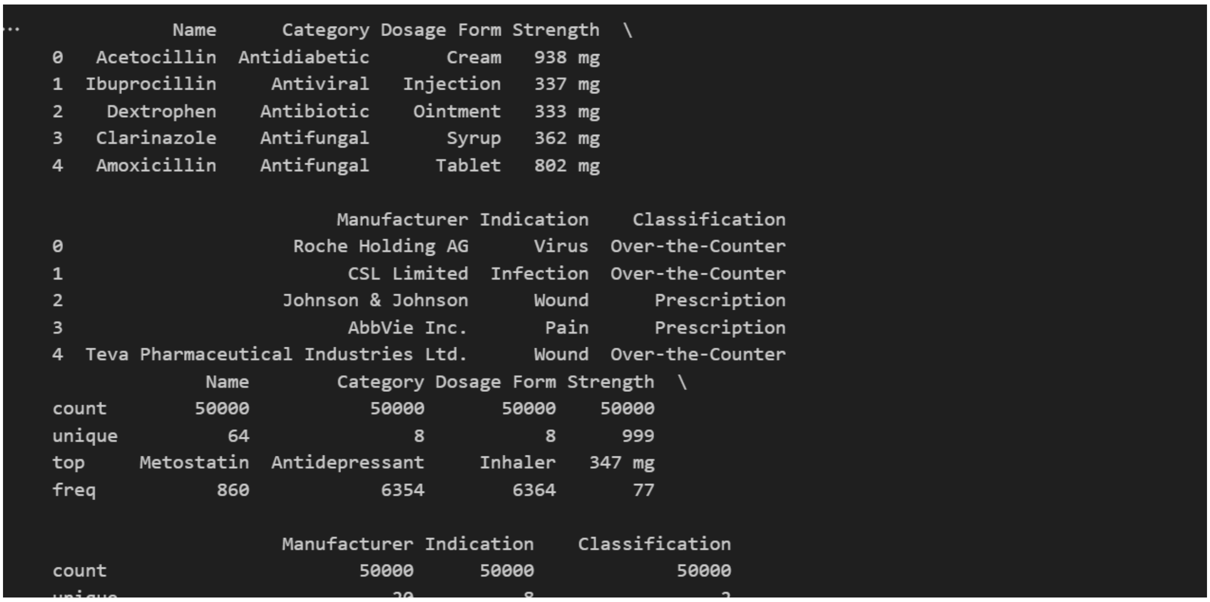


The bar chart shown illustrates the distribution of various drug categories in a dataset. Each category (such as Antidepressant, Analgesic, Antiseptic, etc.) is represented on the x-axis, while the count of drugs in each category is shown on the y-axis.

Based on the chart, it appears that the dataset contains roughly an equal number of entries for each drug category, indicating a balanced distribution. This suggests that the dataset provides an equal representation of different types of drugs, which can be beneficial when conducting analyses or developing models related to drug classification or comparison.



The box plot shown illustrates the distribution of various drug categories in a dataset. Each category (such as Antidepressant, Analgesic, Antiseptic, etc.) is represented on the x-axis, while the count of drugs in each category is shown on the y-axis.



```
...             Name      Category Dosage Form Strength  \
    0   Acetocillin  Antidiabetic       Cream   938 mg   
    1  Ibuprocillin     Antiviral   Injection   337 mg   
    2    Dextrophen    Antibiotic    Ointment   333 mg   
    3   Clarinazole    Antifungal       Syrup   362 mg   
    4   Amoxicillin    Antifungal      Tablet   802 mg   

                          Manufacturer Indication    Classification  
    0                   Roche Holding AG      Virus  Over-the-Counter  
    1                        CSL Limited  Infection  Over-the-Counter  
    2                  Johnson & Johnson      Wound      Prescription  
    3                        AbbVie Inc.       Pain      Prescription  
    4  Teva Pharmaceutical Industries Ltd.      Wound  Over-the-Counter  
                  Name        Category Dosage Form Strength  \
    count        50000           50000       50000    50000   
    unique          64               8           8      999   
    top     Metostatin  Antidepressant     Inhaler   347 mg   
    freq           860            6354        6364       77   

                      Manufacturer Indication    Classification  
    count                    50000      50000             50000  
    unique                      20          8                 2  
```

## 13. MARKET MIX:

- **Segmentation Criteria**:
  1. **Demographics**: Analyze the drug categories based on their intended user demographics (age, gender, health conditions).
  2. **Geography**: If geographical data were available, segmentation could include different regions and their preferences for certain drug types.

- **13.2 Drug Categories**:
  1. The dataset includes multiple drug categories (e.g., Antibiotic, Antifungal, Antipyretic). Segmenting by category helps identify market needs and preferences for specific types of drugs.
  2. Understanding which categories have the highest demand can guide marketing strategies.

- **Strength of Drugs**:
  1. Drug strength can significantly influence prescribing habits and patient adherence. Segmentation based on strength can highlight differences in user preferences or requirements.
  2. Analyzing the average strength of drugs within each segment can provide insights into market positioning.

- **Dosage Forms**:
  1. Different dosage forms (e.g., tablet, syrup, injectable) can affect consumer choice. Segmentation based on dosage forms helps tailor marketing efforts and distribution strategies.

2. Identifying which forms are most popular in specific segments can assist in inventory management and production planning.

- **Manufacturer and Brand Influence**:
  1. If the dataset includes manufacturer information, this can be crucial for segment analysis. Brand loyalty and reputation often influence consumer choices.
  2. Analyzing performance by manufacturer could reveal trends in market share and competitive positioning.

- **Indications and Classifications**:
  1. Segmenting by the indication for which the drug is prescribed (e.g., treatment of infections, pain relief) can lead to targeted marketing and educational campaigns for healthcare professionals.
  2. Classifications based on therapeutic areas (e.g., cardiovascular, respiratory) can provide insights into the competitive landscape.

- **Cluster Analysis Results**:
  1. Using clustering algorithms (like K-Means) to identify distinct market segments helps in understanding consumer behavior.
  2. Interpreting cluster characteristics (e.g., average strength, preferred dosage forms) can inform product development and marketing strategies.

- **Targeted Marketing Strategies**:
  1. Based on the segmentation analysis, marketing strategies can be tailored to address the specific needs and preferences of different segments, optimizing advertising efforts and resource allocation.
  2. Developing unique value propositions for each segment can enhance customer engagement and satisfaction.

- **Potential for Product Development**:
  1. Insights from segmentation can drive innovation in product development by identifying unmet needs in specific segments.
  2. New products or variations of existing products can be created to cater to the preferences of distinct market segments.

- **Performance Monitoring**:
  1. Establishing key performance indicators (KPIs) for each segment can help track the effectiveness of marketing strategies and product offerings over time.
  2. Regularly reviewing segment performance can inform adjustments in strategy and product development.

## 14. Business Modelling

Our product will benefit both healthcare practitioners and individual users by providing fast and accurate disease predictions based on symptoms. The business concept for this system is based on a multi-tiered approach that maintains profitability while providing important health services to users.

**14.1 Product Description:**
Patients or doctors can easily download the app from app stores (Google Play, Apple App Store) or access it via the web. They begin by registering and creating an account, providing basic information and agreeing to terms of use and privacy policies. Users can opt for free access with limited features or choose to subscribe for advanced features such as unlimited disease predictions, detailed reports, and early alerts. Subsequently, individual users can enter their symptoms onto the app at any moment, and the AI will instantly analyse the data to recommend potential conditions. For healthcare practitioners, the app helps doctors during consultations by providing real-time predictions based on patient symptoms, improving diagnosis accuracy. The app is available on a pay-per-use or subscription basis, providing flexibility for occasional users as well as regular access to healthcare institutions. This makes it affordable and scalable, offering value to individuals and medical professionals alike.

This app is intended for a wide variety of users, including:

- **Hospitals and clinics:** Doctors and healthcare professionals can use the system to increase diagnostic accuracy and timeliness.
- **Telemedicine Platforms:** Companies that provide online consultations can include the tool to help virtual doctors make accurate, real-time forecasts.
- **Individual Patients:** People can utilize the system to acquire a general understanding of their condition before going to the doctor, saving time and money.

**14.3 Revenue Streams:**

Our business approach generates cash from three primary sources:

**14.3.3 Subscription Model:**

- **Individual Subscription:** Patients pay a monthly or annual price to receive limitless predictions. For example, people may pay Rs. 500 per month for unlimited symptom checks.
- **Institutional Subscription:** Hospitals or clinics can use the service for a monthly or yearly charge (e.g., Rs. 10,000 per month) to integrate it into their diagnostic procedures.

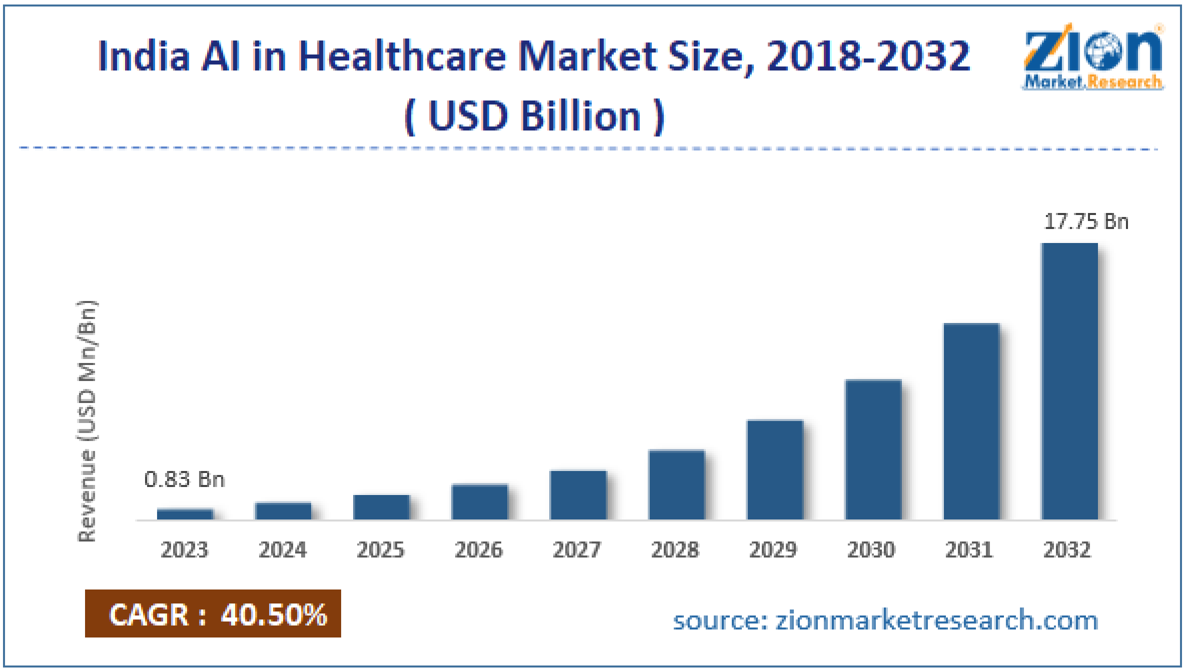
**14.3.4 License Model:**

- **Annual Licensing for Healthcare Providers:** Larger organizations, such as hospitals, can purchase a yearly license to incorporate the technology into their workflow. For example, a hospital would pay Rs. 1,00,000 per year to use the system across all departments.

**14.3.5 Pay-per-Use Model**:

- **One-time Use for Individuals or Small Clinics**: For users who don’t need regular access, the system can charge a fee per prediction request (e.g., Rs. 200 per use). This is useful for patients who may only need the service occasionally or small clinics that don't want to commit to a subscription.

## 15. Financial Modelling



The bar chart provided by Zion Market Research illustrates the projected India AI in Healthcare Market Size from 2018 to 2032, measured in USD billions, with a Compound Annual Growth Rate (CAGR) of 40.50%. This rapid growth demonstrates a substantial increase in AI adoption across the healthcare sector, including for AI-driven disease prediction apps.

Starting Point (2023): The Indian AI healthcare market size in 2023 is $0.83 billion, indicating the early stages of AI-based healthcare tools like disease prediction apps.
Significant Growth by 2032: By 2032, the market size is projected to reach $17.75 billion, highlighting massive growth potential for businesses offering AI healthcare solutions.
CAGR of 40.50%: A CAGR of 40.50% suggests exponential growth in adopting AI technologies in healthcare, which will drive the demand for predictive healthcare applications.

### 15.1 Key Variables in the Financial Model:

### 15.1.1 Unit Price per Subscription:

- **Monthly Subscription (Individuals)**: Rs. 500 per month.
- **Annual Subscription (Individuals)**: Rs. 5,000 per year (discounted from Rs. 6,000).
- **Monthly Subscription (Institutions)**: Rs. 10,000 per month.
- **Annual Subscription (Institutions)**: Rs. 1,00,000 per year.

- **Pay-per-use**: Rs. 200 per use for occasional users or smaller clinics.

### 15.1.2 Fixed Costs:

- Monthly operating costs include server maintenance, development, staff salaries, customer support, and marketing. Assume Rs. 20,000 as the fixed monthly cost.

### 15.1.3 Sales/Subscriptions (denoted as xxx):

- $x\_monthly\_ind: Number\ of\ individual\ monthly\ subscriptions.$
- $x\_monthly\_inst: Number\ of\ institutional\ monthly\ subscriptions.$
- $x\_yearly\_ind: Number\ of\ individual\ yearly\ subscriptions.$
- $x\_yearly\_inst: Number\ of\ institutional\ yearly\ subscriptions.$
- $x\_pay\_per\_use: Number\ of\ pay - per - use\ requests.$

### 15.1.4 Revenue (denoted as y):

- Total income generated from subscriptions, licensing, and pay-per-use services.

## 15.2 Financial Equations for Different Subscription Types

### 15.2.1 Monthly Subscription Revenue:

$$Y\_monthly = 500\ x\_monthly\_ind + 10000\ x\_monthly\_inst - Fixed\ Costs$$

Where:

- x_monthly_ind = number of individual monthly subscriptions.
- x_monthly_inst = number of institutional monthly subscriptions.
- Fixed costs = Rs. 20,000 per month.

### 15.2.2 Yearly Subscription Revenue:

$$Y\_yearly = 5000\ x\_yearly\_ind + 100000\ x\_yearly\_inst - Fixed\ costs$$

Where:

- x_yearly_ind= number of individual yearly subscriptions.
- x_yearly_inst = number of institutional yearly subscriptions.

### 15.2.3 License Revenue:

$$Y\_license = 100000\ x$$

Where:

- x_license = number of healthcare institutions purchasing the yearly license for their operations.

### 15.2.4 Pay-per-Use Revenue:

$$Y\_pay\_per\_use = 200\ x\_pay\_per\_use - Fixed\ Costs$$

- x_pay_per_use = number of pay-per-use requests.

### 15.2.5 Formula for Overall Total Revenue (Yearly):

The overall yearly revenue can be calculated as:

$$Y\_total\_yearly = y\_monthly\_total \times 12 + y\_yearly\_total + y\_pay\_per\_use$$

Where:

- y_monthly_total = Total monthly revenue.
- y_yearly_total= Total yearly subscription revenue.
- y_pay_per_use_total= Total revenue from pay-per-use requests.

## 15.3 Example Calculation

Let's break down the calculation assuming the following:

Assumptions for Yearly Calculation:

- 200 individual monthly subscriptions.
- 5 institutional monthly subscriptions.
- 1,000 pay-per-use requests per month.
- 100 individual yearly subscriptions.
- 3 institutional yearly subscriptions.

### Step 1: Monthly Revenue Calculation

$$Y\_monthly = (500 \times 200) + (10000 \times 5) - 20000$$

Breaking it down:

- Revenue from individual monthly subscriptions = 500×200=100,000 Rs.
- Revenue from institutional monthly subscriptions = 10000×5=50,000 Rs

So, the total monthly revenue is:

$$Y\_monthly = 100{,}000 + 50{,}000 - 20{,}000 = 130{,}000\ Rs$$

### Step 2: Yearly Subscription Revenue Calculation

$$Y\_yearly = (5000 \times 100) + (100000 \times 3)$$

Breaking it down:

- Revenue from individual yearly subscriptions = 5000×100=500,000 Rs
- Revenue from institutional yearly subscriptions = 100000×3=300,000 Rs.

So, the total yearly subscription revenue is:

$$Y\_yearly = 500{,}000 + 300{,}000 = 800{,}000\,Rs.$$

**Step 3: Pay-per-Use Revenue Calculation**

Since we have **1,000 pay-per-use requests per month**, the yearly pay-per-use revenue is:

$$Y\_pay\_per\_use = 200 \times 1000 \times 12 = 2{,}400{,}000\,Rs.$$

**Step 4: Calculate Overall Yearly Revenue**

Now sum the total yearly revenue:

$$Y\_total\_yearly = (130{,}000 \times 12) + 800{,}000 + 2{,}400{,}000$$

Breaking it down:

- Total monthly revenue for 12 months = 130,000×12=1,560,000 Rs.
- Total yearly subscription revenue = 800,000 Rs.
- Total pay-per-use revenue = 2,400,000 Rs.

So, the overall total yearly revenue is:

$$Y\_total\_yearly = 1{,}560{,}000 + 800{,}000 + 2{,}400{,}000 = 4{,}760{,}000\,Rs$$

### 15.4 Final Revenue

The overall total revenue for the whole year, considering monthly subscriptions, yearly subscriptions, and pay-per-use requests, is **Rs. 4,760,000**.

## 16. Conclusion

The financial model for the disease prediction app includes multi-tiered revenue streams based on subscriptions (monthly and yearly), licensing for healthcare providers, and pay-per-use options. By starting with a base of subscriptions and assuming a steady 10% growth rate per month, the app has the potential to scale quickly and reach profitability in the growing AI healthcare market.