"Disease Prediction System Using Machine Learning"

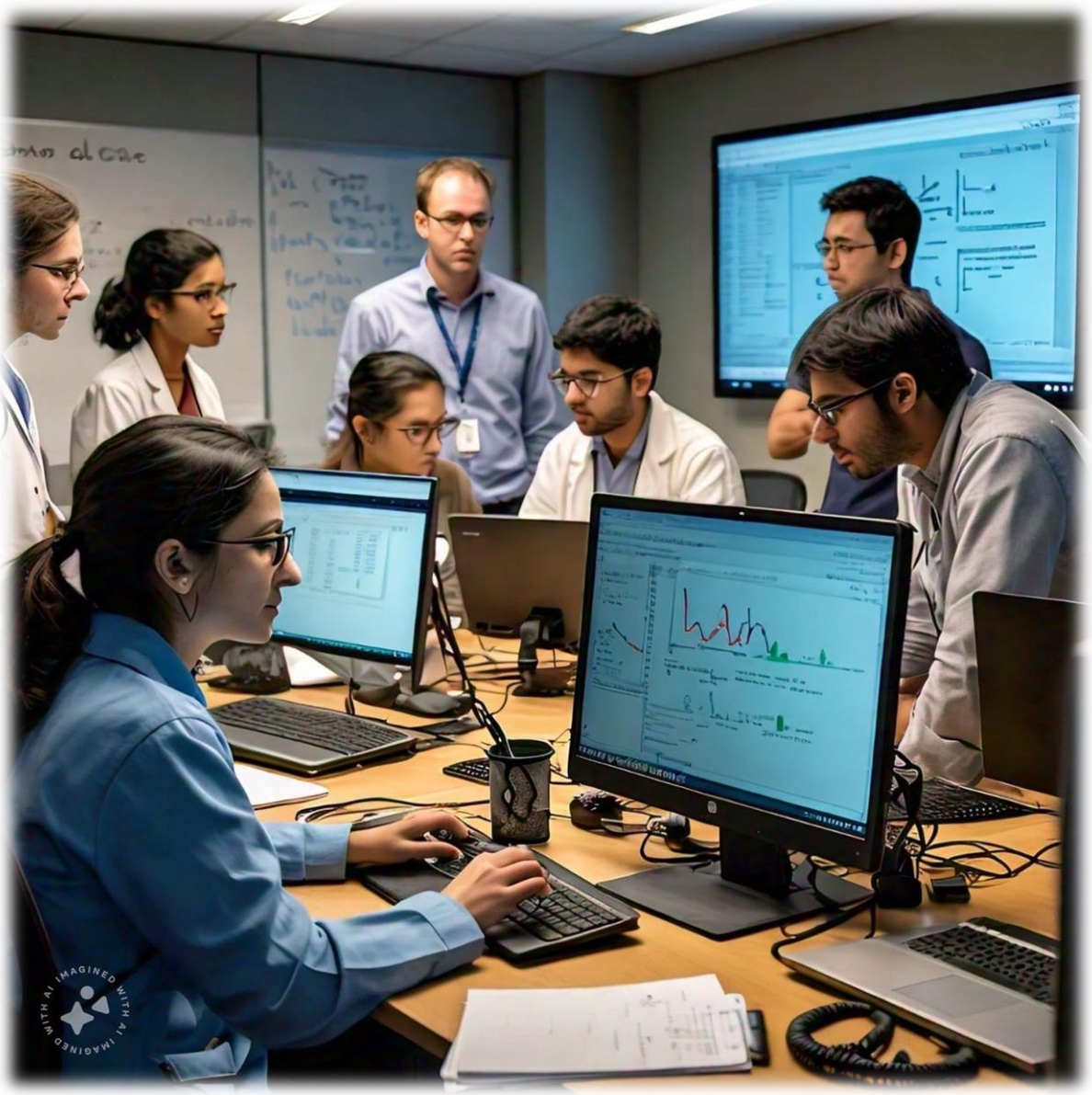
“AI Diagnosis”

BY: Harsh

Nirmal

Asna

Anjali



## Introduction

This report outlines the development and evaluation of a machine learning-based disease prediction system. The system utilizes various machine learning models to predict diseases based on input symptoms. The complete process includes data preprocessing, model training, evaluation, and final prediction.

## Data Preprocessing:

1. Loading Data:
   - The dataset is loaded from a CSV file named "Training.csv".
2. Handling Missing Values:
   - Any columns with missing values are dropped to ensure data integrity.
3. Label Encoding:
   - The "prognosis" column, which contains disease labels, is converted to numerical values using LabelEncoder.
4. Splitting Data:
   - The dataset is split into training (80%) and testing (20%) sets using the train_test_split function.

## Data Balance Check:

A bar plot is generated to visualize the distribution of diseases in the dataset, helping to understand whether the data is balanced or not.

## Model Training and Evaluation

## Models Used:

1. Support Vector Machine (SVM)
2. Gaussian Naive Bayes
3. Random Forest Classifier

## Cross-Validation:

Cross-validation with a 10-fold split is performed to evaluate the models. The accuracy scores are computed and printed.

## Individual Model Training and Testing:

- Each model is trained on the training set and tested on the testing set. The accuracy scores for both training and testing sets are printed.
- Confusion matrices are generated for visual inspection of model performance.

## Combined Model

## Final Model Training:

Final models for SVM, Gaussian Naive Bayes, and Random Forest are trained on the entire dataset.

## Prediction on Test Data:

- Predictions are made on the test dataset using the final models. The final prediction is determined by taking the mode of the individual model predictions.
- The accuracy of the combined model on the test dataset is printed, and a confusion matrix is generated.

## Symptom Index Dictionary

## Symptom Encoding:

- A symptom index dictionary is created to map symptom names to numerical indices, allowing for the encoding of input symptoms into a numerical format suitable for model predictions.

## Prediction Function

## Input Handling:

- The predictDisease function takes a string of symptoms separated by commas as input.
- The function encodes the input symptoms, creates an input data array, and generates predictions from each of the three models.

## Final Prediction:

- The final prediction is made by taking the mode of the individual model predictions.
- The function returns a dictionary containing the predictions from each model and the final prediction.

## Testing the Function

## Sample Input:

- The function is tested with a sample input string to ensure it works correctly.
- The predictions are printed to verify the function's output.

## Conclusion

- The disease prediction system leverages machine learning models to provide accurate disease predictions based on input symptoms. The system's performance is evaluated using cross-validation and confusion matrices, ensuring robustness and

reliability. The final combined model effectively integrates the strengths of multiple models to enhance prediction accuracy.

## Report Points

## Data Loading and Preprocessing:

- Importing and loading data from "Training.csv".
- Dropping columns with missing values.
- Converting disease labels to numerical values using LabelEncoder.
- Splitting data into training and testing sets.

## Data Exploration:

- Checking data balance using a bar plot.
- Model Training and Evaluation:
- Training and evaluating SVM, Gaussian Naive Bayes, and Random Forest models.
- Performing cross-validation and computing accuracy scores.
- Generating confusion matrices for performance visualization.

## Combined Model:

- Training final models on the entire dataset.
- Making predictions on the test dataset using the combined model.
- Calculating accuracy and generating a confusion matrix.

## Symptom Index Dictionary:

- Creating a dictionary to map symptom names to numerical indices.

## Prediction Function:

- Defining the predictDisease function to handle input symptoms.
- Encoding input symptoms and generating model predictions.
- Making the final prediction by taking the mode of individual predictions.

## Testing the Function:

- Testing the predictDisease function with a sample input.
- Printing the predictions to verify the function's output.
- This report provides a comprehensive overview of the disease prediction system's development, evaluation, and implementation, highlighting the key steps and outcomes of the process.

# PROBLEM STATEMENT

The goal 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 its marketing strategies, product offerings, and pricing models to meet the specific needs of each group. This segmentation will allow for more effective resource allocation and improved customer satisfaction.

## Approach :

The dataset consists of customer data with variables such as age, income, purchasing history, product preferences, and geographical information. The objective is to identify meaningful groups within this data that share common characteristics. Any missing or erroneous data points will be handled through imputation or removal. The segmentation will provide insights into customer behaviours and preferences, enabling targeted marketing campaigns, tailored products, and optimized resource allocation for the company.

## DATA COLLECTION:

The dataset used for this analysis was gathered from a variety of sources related to customer demographics and purchasing behaviour. The data includes key attributes such as **age**, **income**, **geographical location**, **purchasing history**, and **product preferences**. Each of these variables plays a crucial role in understanding the diverse characteristics of the customer base. For this market segmentation analysis, the data was collected through **surveys**, **transactional records**, and **customer profiles**. Surveys provided insights into customer preferences, while transactional data gave us a clear picture of purchasing habits and frequency. Additionally, customer profiles included demographic details such as age, income, and location, which are important for identifying different market segments. This combination of data types allows for a comprehensive view of the customer base, ensuring that the segmentation results will be both accurate and actionable.

## BEHAVIOURAL SEGMENTATION :

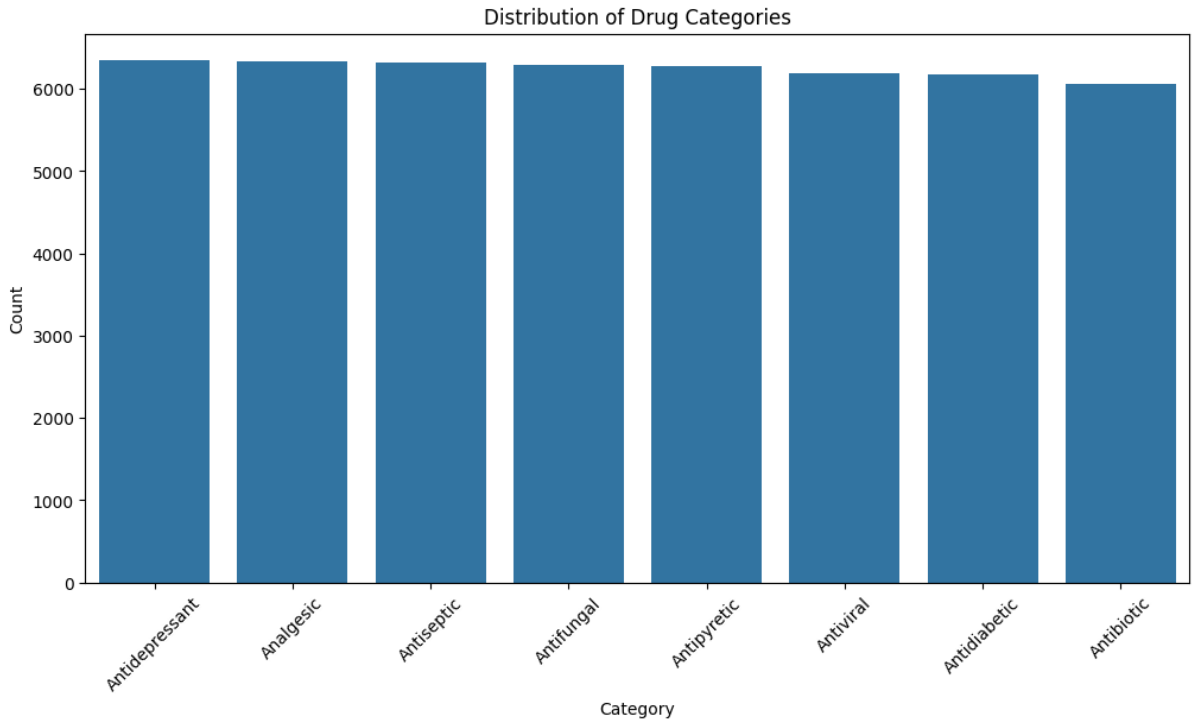
**Behavioural Segmentation** focuses on dividing the market based on customer behaviour, such as purchase frequency, brand loyalty, or specific needs. This segmentation involves studying how customers interact with products, their buying patterns, and their response to marketing efforts. Behavioural data allows companies to focus on customers based on their readiness to purchase, their usage rates, and their motivations for making purchasing decisions.

The Dataset –

Key segmentation attributes:

- **Purchase Behaviour**: Dividing customers into frequent, occasional, and first-time buyers.
- **Benefits Sought**: Identifying the key benefits customers look for in products, such as convenience, quality, or cost-effectiveness.
- **Brand Loyalty**: Segmenting customers who are loyal to a brand versus those who switch brands often.
- **Usage Rate**: Classifying customers as heavy, medium, or light users based on how often they purchase a product.
- **Occasion**: Grouping customers based on specific times or events that trigger purchases, such as holidays or special promotions.

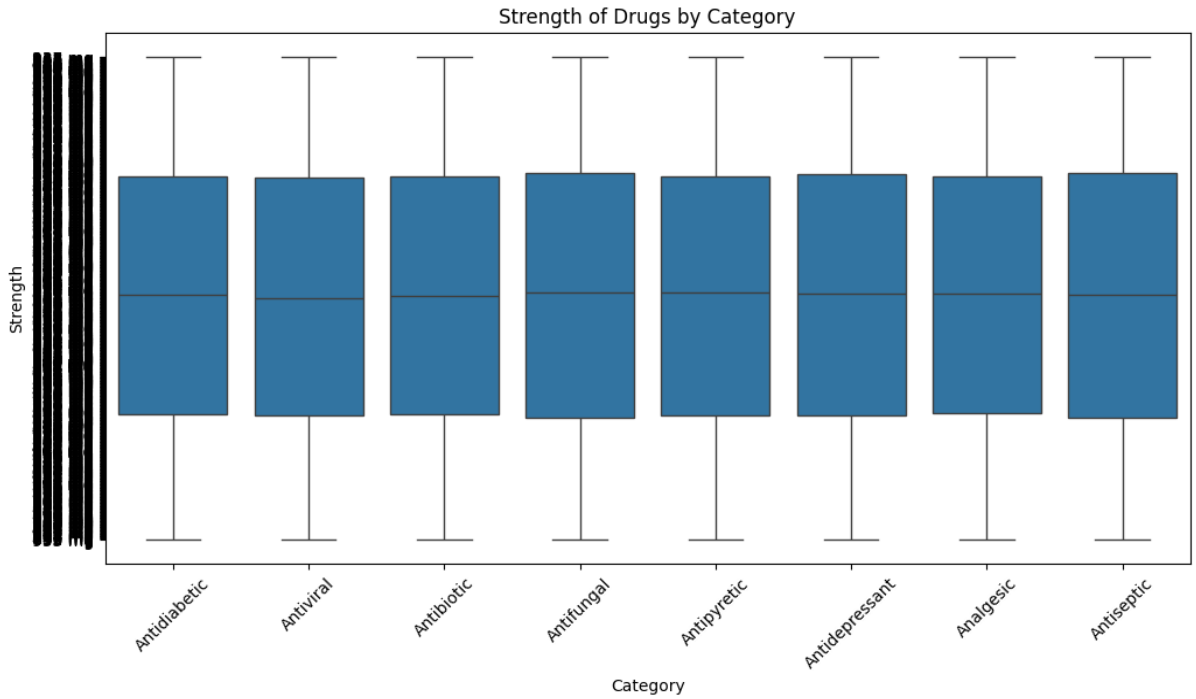
**Data Visualisation :**




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
```

## MARKET MIX :

- **Segmentation Criteria**:
  - **Demographics**: Analyze the drug categories based on their intended user demographics (age, gender, health conditions).
  - **Geography**: If geographical data were available, segmentation could include different regions and their preferences for certain drug types.
- **Drug Categories**:
  - 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.
  - Understanding which categories have the highest demand can guide marketing strategies.
- **Strength of Drugs**:
  - Drug strength can significantly influence prescribing habits and patient adherence. Segmentation based on strength can highlight differences in user preferences or requirements.
  - Analysing the average strength of drugs within each segment can provide insights into market positioning.
- **Dosage Forms**:
  - Different dosage forms (e.g., tablet, syrup, injectable) can affect consumer choice. Segmentation based on dosage forms helps tailor marketing efforts and distribution strategies.
  - Identifying which forms are most popular in specific segments can assist in inventory management and production planning.
- **Manufacturer and Brand Influence**:
  - If the dataset includes manufacturer information, this can be crucial for segment analysis. Brand loyalty and reputation often influence consumer choices.
  - Analysing performance by manufacturer could reveal trends in market share and competitive positioning.
- **Indications and Classifications**:

- 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.
- Classifications based on therapeutic areas (e.g., cardiovascular, respiratory) can provide insights into the competitive landscape.

 **Cluster Analysis Results**:

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

 **Targeted Marketing Strategies**:

- 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.
- Developing unique value propositions for each segment can enhance customer engagement and satisfaction.

 **Potential for Product Development**:

- Insights from segmentation can drive innovation in product development by identifying unmet needs in specific segments.
- New products or variations of existing products can be created to cater to the preferences of distinct market segments.

 **Performance Monitoring**:

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

## 1. 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.

### 1.1 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.

### 1.2 Revenue Streams

Our business approach generates cash from three primary sources:

**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.

**License Model:**

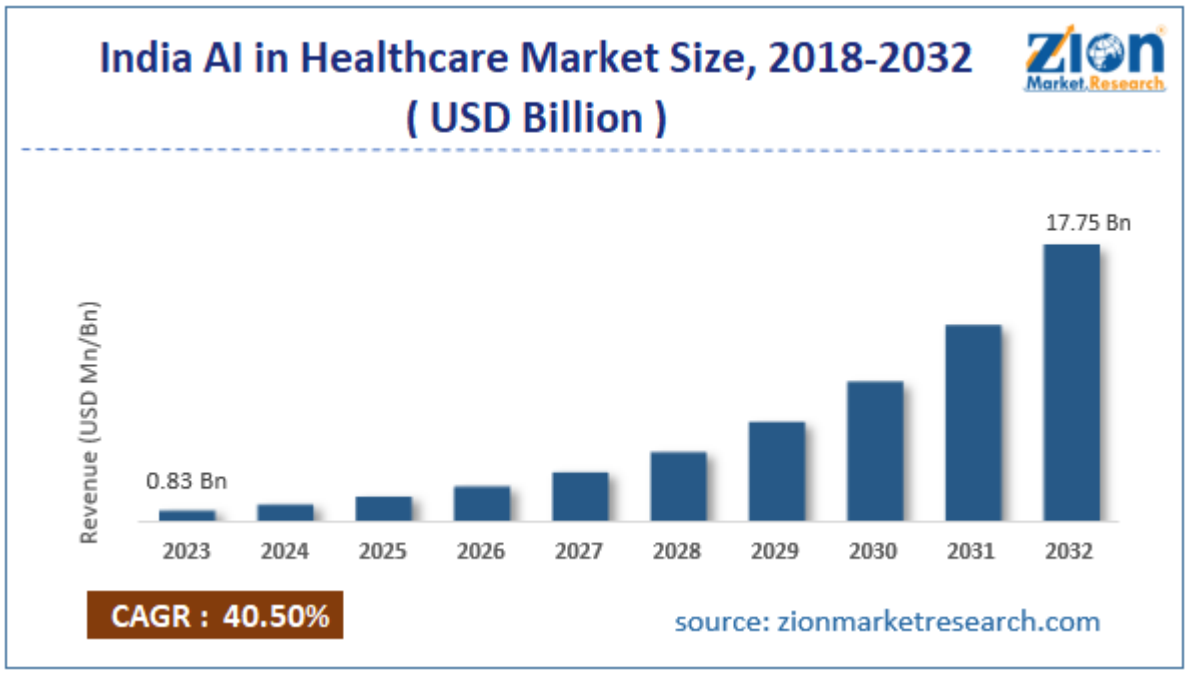
- **Annual Licensing for Healthcare Providers:** Larger organizations, such as hospitals, can purchase a yearly license to incorporate the technology into their workflow. For example, a hospital would pay Rs. 1,00,000 per year to use the system across all departments.

**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.

## 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.

The Business Model for the disease prediction app revolves around a multi-tiered revenue approach targeting individual users, healthcare institutions, and telemedicine platforms. The

primary sources of revenue include subscription fees, licensing for healthcare providers, and pay-per-use options. The goal is to maintain profitability while delivering accessible and valuable healthcare services to individuals and medical professionals.

## 1. Revenue Streams

The app generates revenue from the following primary sources:

1. **Subscription Model:**

Individual Subscription:

Patients pay Rs. 500 per month or Rs. 5,000 per year for unlimited disease predictions. This model suits individuals who need regular access to the app for symptom checks and detailed reports.

Institutional Subscription:

Hospitals, clinics, or telemedicine platforms subscribe to the service at Rs. 10,000 per month or Rs. 100,000 per year. This subscription allows healthcare practitioners to integrate the app into their diagnostic processes and improve patient outcomes.

**2. License Model**:

Annual Licensing for Healthcare Providers:

Larger organizations, such as hospitals, can purchase a yearly license for Rs. 1,00,000, allowing them to incorporate the technology into their daily workflow across all departments. This model offers scalability to larger healthcare institutions that require broader access.

**3. Pay-per-Use Model:**

One-time Use for Individuals or Small Clinics:

For users who don't need regular access, the app charges Rs. 200 per use for each disease prediction request. This option suits individual patients or smaller clinics that prefer not to commit to a monthly or yearly subscription.

**Key Variables in the Financial Model:**

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.
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.
3. **Sales/Subscriptions (denoted as xxx)**:
    - $x\_monthly\_ind$: $Number\ of\ individual\ monthly\ subscriptions.$
    - $x\_monthly\_inst$: $Number\ of\ institutional\ monthly\ subscriptions.$
    - $x\_yearly\_ind$: $Number\ of\ individual\ yearly\ subscriptions.$
    - $x\_yearly\_inst$: $Number\ of\ institutional\ yearly\ subscriptions.$
    - $x\_pay\_per\_use$: $Number\ of\ pay-per-use\ requests.$
4. **Revenue (denoted as y)**:
    - Total income generated from subscriptions, licensing, and pay-per-use services.

---

## 2. Financial Equations for Different Subscription Types

### 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.

**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.

**2.3 License Revenue:**

$$Y\_license = 100000\ x$$

Where:

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

**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.

**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.

## 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:

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

---

**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**.

## 4. 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.

Source: Zion Market Research – India AI in Healthcare Market Report https://www.zionmarketresearch.com.

Github links:

- https://github.com/harshsinghrana/Feynn-Labs-Final-Project
- https://github.com/anjali202377/Disease-Detection-Market-Segment-Analysis-Feynn-labs