HEALTHSAGE: AI-POWERED HEALTH PREDICTION AND RISK ASSESSMENT APP FOR MEDICAL LAB COMPANIES

Febin Francis

20/09/2024

Abstract

HealthSage is an AI-powered health prediction and risk assessment app designed for medical labs in India. The objective is to provide users with personalized health insights by analyzing lab test results, medical history, lifestyle data, and genetic information. By using machine learning models, the app predicts future health risks and offers preventive measures, including dietary recommendations, exercise routines, and specialist referrals. The report covers the market need, product design, technical specifications, and business model. The results show that HealthSage fills a critical gap in preventive healthcare, offering a comprehensive approach to managing future health risks.

1.0 Problem Statement:

India's healthcare system is still largely reactive, with patients seeking medical attention only after symptoms manifest. However, there is a growing need for preventive healthcare solutions that can help predict future health risks based on lab results, medical history, lifestyle factors, and genetics. Despite the availability of lab results, most individuals lack access to an easy-to-understand analysis of their data that can link these results with personalized health insights.

Traditional lab reports offer diagnostic results but provide little guidance on what these results mean in relation to one's overall health. With increasing lifestyle diseases like diabetes, heart disease, and obesity in India, there is a need for a comprehensive platform that not only predicts future health risks but also provides actionable insights for preventing potential health problems.

This paper proposes an AI-powered health prediction app designed specifically for the Indian healthcare market. This app will integrate lab test results, family medical history, lifestyle habits, and genetic predispositions to provide personalized health risk assessments, preventive recommendations, and specialist referrals.

2.0 Market/Customer/Business Need Assessment:

2.1 Market Need:

India is experiencing a surge in lifestyle-related health issues, making preventive healthcare a priority for millions. However, interpreting medical test results and understanding how they relate to lifestyle habits and potential future health risks is difficult for many individuals. Predictive healthcare, supported by AI, is becoming essential to combat this issue. This app targets the large and diverse Indian market, which is increasingly adopting digital health tools.

2.2 Customer Need:

Patients in India are more health-conscious and tech-savvy, looking for personalized health solutions that give them insights into their future risks and actionable ways to mitigate them. Many individuals, especially those in urban areas, actively seek information about their health, often using apps and devices to track fitness, diet, and overall well-being. However, they face a gap in terms of interpreting lab results in a meaningful and preventive context. This app aims to bridge that gap by making complex lab data easy to understand while providing proactive health advice.

2.3 Business Need:

Medical labs and diagnostic centers are under constant pressure to differentiate themselves in a highly competitive market. Offering AI-driven predictive health assessments can be a key differentiator. This app allows medical labs to provide an added layer of value to their customers, turning lab results into comprehensive health insights, which helps build customer loyalty and generate recurring revenue through premium services.

3.0 Target Specifications and Characterization:

3.1 Primary Customers:

Medical labs in India are looking to provide an additional layer of predictive insights to their patients alongside lab test results.

3.2 End Users:

- Health-conscious individuals looking for proactive healthcare insights.
- People with a family history of lifestyle diseases (e.g., diabetes, heart disease).
- Users who frequently get lab tests for monitoring chronic conditions.
- Individuals using wearable fitness trackers who want deeper integration of their health data with lab results.

3.3 Characteristics:

- Provides a comprehensive health profile using AI to predict future risks based on lab results, lifestyle, and genetics.
- Offers preventive measures, such as diet and exercise suggestions, and specialist referrals for high-risk conditions.
- Subscription models offer personalized disease-specific reports and continuous monitoring.

4.0 External Search:

- **AI in Healthcare**: Machine learning models have been successful in predicting specific health outcomes (e.g., risk of heart disease).
- **Personalized Medicine**: Studies have shown the benefits of using individual health data, including genetic, lifestyle, and clinical data, to tailor health interventions.
- **Existing Apps:** Applications like MyFitnessPal and Google Fit, which track fitness and lifestyle but don't offer predictive health insights based on lab data.
- **Market reports**: Increasing demand for AI in healthcare solutions and preventive health apps.

5.0 Benchmarking Alternate Products:

- **Lab Informer**:
  Lab Informer focuses on simplifying the interpretation of lab reports by offering

explanations for each test result. It is excellent for users looking to better understand their diagnostic results but lacks in-depth predictive analysis or personalized preventive recommendations based on lifestyle and genetic factors.

- **Docus**:
  Docus offers a lab-test interpretation tool that uses AI to predict possible medical conditions based on test results. While it provides valuable diagnostic insights, its focus is primarily on identifying potential current health issues rather than a comprehensive approach to predicting future risks or recommending preventive strategies.
- **HealthifyMe**:
  A well-known fitness and nutrition app in India, HealthifyMe provides users with diet and exercise plans but does not integrate lab data or AI-powered health risk assessments.
- **Apple Health**:
  Strengths: Robust data aggregation from fitness apps, wearables, and healthcare data.
  Weaknesses: Lacks a machine learning model for predicting future health issues and doesn't offer genetic testing integration.

6.0 Applicable Indian Laws and Patents:

- **Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011**:
  The app must comply with India's regulations regarding the protection of sensitive personal health data.
- **Personal Data Protection Bill, 2019** (pending):
  This bill, once enacted, will introduce stricter controls over the handling of personal health data. The app must be built with a privacy-first approach, ensuring that users' medical data is processed securely and in compliance with data protection laws.
- **Medical Device Rules, 2017**:
  If the app is considered a medical device under the Indian regulatory framework, it will need to comply with these rules, particularly in terms of safety and effectiveness.
- **Patents in AI Healthcare**:
  The algorithms that power the health risk prediction models can potentially be patented under the **Indian Patents Act, 1970**, provided they meet the criteria for technical innovation and novelty.

7.0 Applicable Regulations:

- **Data Privacy**:
  The app must comply with India's **Information Technology Act, 2000**, **IT(Amendment) Act, 2008** and other regulations around data privacy, particularly those related to the storage and sharing of health data.

- **Medical Council of India (MCI) Regulations**:
  The app should ensure that any medical advice provided, especially when referring users to specialists, adheres to the guidelines set by MCI. AI predictions must not replace licensed healthcare providers.
- **Consumer Protection Act, 2019:**
  Consumer Rights: Ensure the app protects consumer rights, including transparency in pricing, clear terms of service, and mechanisms for addressing consumer complaints.
  Grievance Redressal: Establish a system for addressing consumer grievances, including the appointment of a grievance officer.
- **Payment and Settlement Systems Act, 2007:**
  Payment Gateways: Ensure that the app's payment gateway integration complies with regulations governing electronic payments and settlements.
  Secure Transactions: Implement secure payment processes and protect user financial data.
- **Goods and Services Tax (GST):**
  Tax Compliance: Register for GST if the annual turnover exceeds the threshold limit and ensure compliance with tax filing requirements.
  Invoicing: Issue GST-compliant invoices for subscription fees, delivery charges, and any other payments received through the app.

8.0 Applicable Constraints:

- **Budget Constraints**:
  The development of this app will require substantial upfront investment in AI model development, partnerships with medical labs, and legal compliance with Indian regulations on health data.
- **Data Sensitivity**:
  Handling sensitive medical data, including lab results, family medical histories, and genetic data, requires compliance with stringent data protection regulations, and any breach could harm user trust.
- **Trust & Adoption**:
  Convincing users that AI-powered health predictions are reliable, accurate, and privacy-focused can be a significant challenge in the Indian market, where trust in data handling remains a concern.

9.0 Business Model:

9.1 Freemium Model:

Basic features, such as lab result interpretation and general health scores, will be free to users.

Premium services, including personalized risk assessments, preventive measures, and access to specialist recommendations, will be available through subscription tiers.

9.2 Premium Packages:

Users can choose from different packages offering more detailed analysis of specific diseases (e.g., heart disease, diabetes) or additional lab tests. These packages increase in price with the number of tests required and the complexity of the insights provided. Higher-tier packages also refer users to specialized doctors for further consultation.

9.3 Subscription Model:

Users pay monthly or annually for ongoing access to premium features like quarterly health checks, updated risk assessments, and personalized health plans.

9.4 Commissions from Referrals:

- **Doctor Referrals:**
  - Users identified as high-risk for specific health conditions can be referred to nearby specialists such as cardiologists, endocrinologists, or nutritionists. HealthSage will partner with these healthcare providers, earning a referral commission for each successful consultation or follow-up booked through the app.
  - This will create a symbiotic relationship where doctors receive new patients while HealthSage earns a fee for providing valuable referrals.
- **Lab Referrals:**
  - The app can partner with diagnostic labs for users who need specific tests beyond their routine checkups. When users book lab tests through HealthSage, the app earns a commission from the labs based on the volume and type of tests ordered.
  - This service benefits labs by driving additional revenue and ensuring higher customer engagement while improving the user's health experience through streamlined test bookings.
- **Fitness Center and Gym Referrals:**
  - HealthSage will collaborate with fitness centers, gyms, and personal trainers. Users with health goals such as weight management, improved cardiovascular health, or increased fitness will be referred to partner fitness centers. The app will earn referral commissions for every subscription or personal training package booked through the platform.
  - Additionally, this encourages users to stay active and healthy, directly aligning with the app's preventive health focus.
- **Diet and Nutrition Subscriptions:**
  - HealthSage can partner with companies that offer health-focused meal delivery services or nutrition programs. Users who need dietary changes (e.g., for

managing diabetes or high cholesterol) can subscribe to these services through the app, generating referral income for HealthSage.
    - This partnership enhances user outcomes by offering convenient access to tailored nutrition plans, ensuring adherence to dietary recommendations.

9.5 Scalability of Referral Commissions:

- **Healthcare Ecosystem Partnerships:** As the app grows, partnerships with pharmacies, mental health professionals, and holistic health providers (e.g., yoga centers, meditation coaches) can further expand the revenue potential through referral models.
- **User Experience Improvement:** By streamlining access to these third-party services, HealthSage offers a seamless experience that integrates medical advice with action steps, increasing user satisfaction and retention, which, in turn, increases referral revenue.

10.0 Concept Generation:

The concept was generated by identifying a gap in the health tech industry—most existing tools either focus on fitness tracking or offer isolated genetic insights, but few provide an integrated view of lab results, lifestyle, and genetic data for predictive health. With the growing focus on personalized and preventive healthcare, it became evident that combining these diverse datasets would offer more valuable insights to users.

The idea behind this app emerged from the growing need for a comprehensive AI-driven platform that can integrate various sources of health data to predict future risks. The current apps like Lab Informer and Docus focus mainly on lab result interpretation and diagnostic analysis. This new app will not only provide easy-to-understand lab results but also integrate long-term health risk predictions and preventive care recommendations based on a holistic view of the user's health data.

11.0 Concept Development:

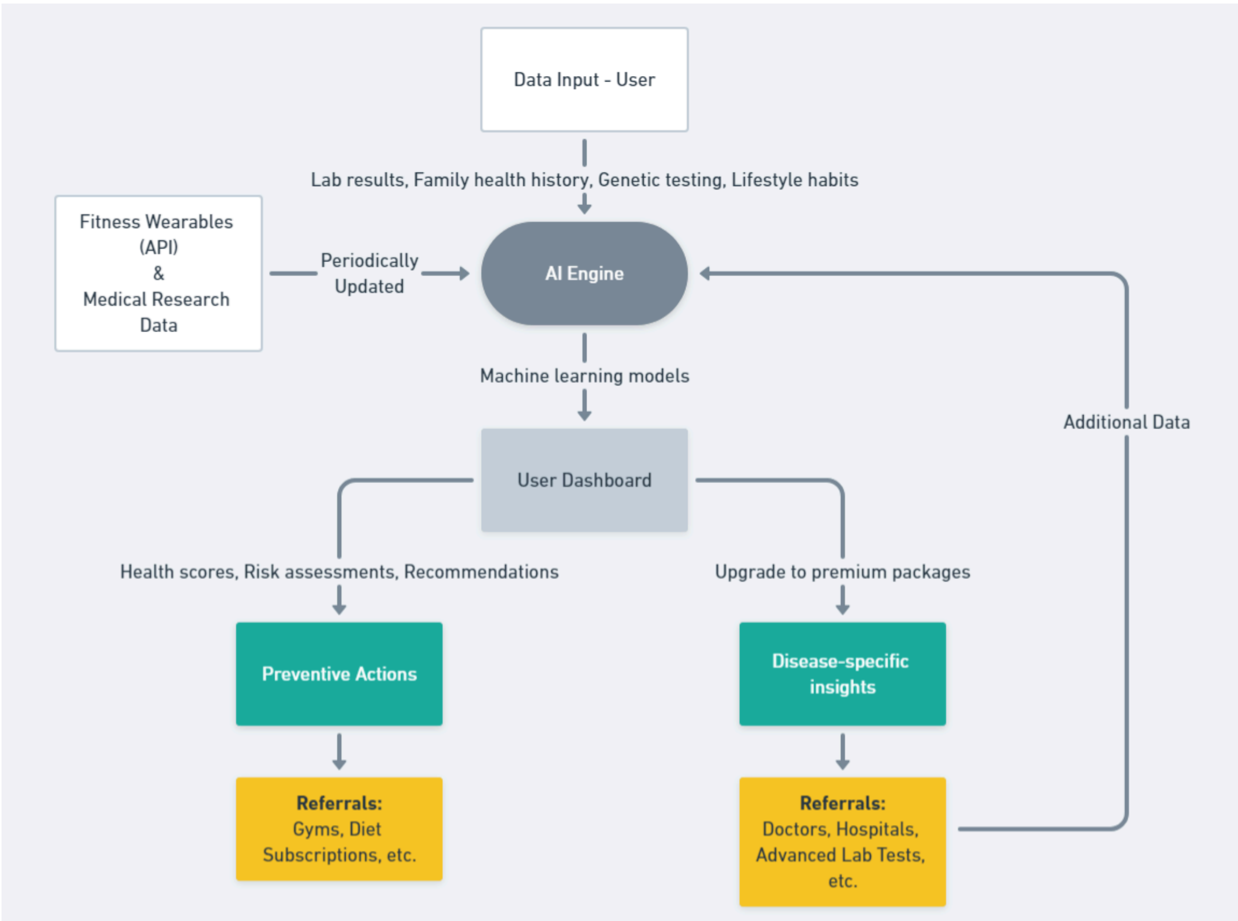
The app will gather health data from multiple sources, such as lab tests, medical history, lifestyle habits, and genetic data, and use AI-powered models to predict future health risks. It will provide users with actionable preventive strategies like dietary recommendations, workout plans, and stress management techniques. The app will also recommend relevant specialists based on identified health risks and encourage users to take proactive measures.

12.0 Final Product Prototype (Abstract) with Schematic Diagram:

Prototype Overview:
An AI-driven health prediction app tailored to the Indian healthcare market, integrating medical

history, lab results, genetic data, and lifestyle habits. The app provides users with personalized health scores, risk assessments, preventive care recommendations, and specialist referrals.

Schematic Diagram:



13.0 Product Details:

13.1 How It Works:

HealthSage is designed to integrate a broad range of user data to provide personalized health risk assessments and recommendations. Here's how the system operates:

- **Data Collection:**
  - Users input their medical history, lab test results, family health history, lifestyle habits, and wearable device data into the app.
  - The app also syncs with third-party health tracking platforms like Fitbit or Apple Health, integrating live data such as daily activity levels, sleep patterns, and heart rate.

- **Data Processing & Analysis:**
    - Once user data is collected, the app uses machine learning models to analyze and generate personalized health insights. These insights include predicting potential future health issues and calculating a **Health Risk Score**, summarizing the user's overall health.
- **Recommendation Engine:**
    - Based on the analysis, HealthSage provides tailored preventive recommendations such as dietary changes, exercise routines, and medical checkups. The app tracks user adherence to these suggestions and updates the recommendations as their health data evolves.
- **Feedback Loop & Continuous Learning:**
    - The app incorporates user feedback and updates its machine learning models over time. For example, if a user follows the dietary advice and improves their health, the app adjusts its future predictions and recommendations accordingly.

13.2 Data Sources:

HealthSage collects and integrates data from a variety of sources, including:

- **User-Provided Data:**
    - Lab test results (e.g., blood tests, glucose levels, cholesterol)
    - Medical history (e.g., previous diagnoses, surgeries, family health history)
    - Lifestyle habits (e.g., diet, smoking, alcohol consumption, exercise routine)
    - Genetic testing results (23andMe, AncestryDNA)
- **Wearable Devices:**
    - Data from devices like Fitbit, Apple Watch, and other fitness trackers. Metrics include heart rate, physical activity, sleep duration, and calories burned.
- **External Health Apps:**
    - Integration with health tracking platforms (e.g., Google Fit, Apple Health) to aggregate user data across different systems.
- **Public Health Databases:**
    - Epidemiological data, health studies, and global databases for reference and benchmarking health trends (optional for specific cases).

13.3 Algorithms, Frameworks, Software Needed:

HealthSage relies on several machine learning models and software tools to deliver accurate predictions and recommendations. Here are the key components:

- **Machine Learning Models:**

    - **Machine Learning Models** like **Logistic Regression, Random Forest** and **Gradient Boosting Machines (GBM)** for health risk prediction.
    - **Neural Networks** (optional) for genetic data analysis.
    - **Clustering Algorithms** (e.g., K-means) for user segmentation based on health profiles.
    - **Natural Language Processing (NLP)** for analyzing family health history and feedback in free-text format.
- **Frameworks and Libraries:**
    - **Python** for data processing and machine learning model development.
    - **TensorFlow** or **PyTorch** for deep learning (Neural Networks).
    - **Scikit-learn** for traditional machine learning algorithms.
    - **Pandas**, **NumPy** for data manipulation.
    - **Matplotlib**, **Plotly, Seaborn** for visualizing health data and predictions.
    - **Flask** or **Django** for backend web framework.
- **Cloud Infrastructure:**
    - **Amazon Web Services (AWS)**, **Google Cloud**, or **Microsoft Azure** for scalable cloud computing and data storage.
    - **Docker** for containerization of the app.
    - **Kubernetes** for orchestration and scaling.
- **Database Management:**
    - **PostgreSQL**, **MySQL**, or **MongoDB** for storing user data and health metrics.
- **API Integration:**
    - Integration with health platforms and wearable devices using **OAuth** or proprietary APIs.

13.4 Team Required to Develop HealthSage:

- **Product Manager (1):**
    - Responsible for overseeing the entire product development process, ensuring alignment with business goals, and managing timelines and budgets.
- **AI/ML Engineers (2-3):**
    - Focus on building and fine-tuning machine learning models for health prediction, including data preprocessing, model training, and deployment.
- **Full-Stack Developers (2-3):**
    - Build both the front-end and back-end of the application, integrate APIs, and ensure smooth interaction between the user interface and the machine learning engine.
- **Data Scientists (2):**
    - Analyze and interpret user data, build predictive models, and provide data-driven insights for feature development.
- **UI/UX Designer (1):**

    - Design a user-friendly interface, focusing on easy navigation, effective visualization of health metrics, and a seamless user experience.
- **DevOps Engineer (1):**
    - Ensure smooth deployment of the application on cloud platforms, manage server infrastructure, and handle containerization and orchestration using tools like Docker and Kubernetes.
- **Mobile App Developers (2):**
    - Develop native or cross-platform mobile apps (Android and iOS), ensuring compatibility with mobile health tracking features and wearable devices.
- **Healthcare Consultant/Advisor (1):**
    - Provide domain expertise to ensure compliance with healthcare regulations, accuracy in health risk assessments, and relevance in recommendations.
- **QA Engineers (1-2):**
    - Test the application to ensure all features work as intended and the app performs well under different user scenarios.
- **Compliance & Legal Expert (1):**
    - Ensure compliance with health data privacy laws like HIPAA, GDPR, and Indian healthcare regulations.

13.5 What Does It Cost?

- **Total Team Salary Estimate (for 6-12 months development cycle):** ₹1.2 crore to ₹1.8 crore
- **Technology and Tools:**
    - **Cloud Hosting (AWS, Google Cloud, etc.):** ₹10-15 lakhs annually (depending on data volume and user base)
    - **Software Licenses and Subscriptions:** ₹3-5 lakhs annually for development tools, APIs, etc.
    - **Wearable Integration (API Fees):** ₹2-5 lakhs annually (for API access to platforms like Fitbit, Apple Health, etc.)
    - **AI/ML Model Training Costs (Cloud Compute):** ₹10-20 lakhs (depending on computational needs)
- **Marketing & Launch:**
    - **App Launch and Marketing Campaign:** ₹10-15 lakhs (initial phase)
    - **Ongoing User Acquisition and Retention:** ₹10-25 lakhs annually (depending on region, platform, and user base)

13.6 Estimated Total Cost:

- **Development (1st year):**
  ₹1.5 crore to ₹2.5 crore
  (Includes team salaries, cloud hosting, software tools, and API integrations)
- **Ongoing Operational Costs (subsequent years):**
  ₹50 lakhs to ₹75 lakhs annually
  (Includes cloud hosting, user support, marketing, and product updates)

15.0 Code Implementation/ Small Scale Validation

The following Python code is designed to process and analyze health data from the **NHANES (National Health and Nutrition Examination Survey) 2005-2006 datasets**. The primary goal of the analysis is to predict key health outcomes based on various biometric and lifestyle factors.

Note: The accuracy of the models are low due to lack of additional features and in depth data pre processing. Model accuracy can be increased further.

Data Source:
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2005

The code starts in the next page:

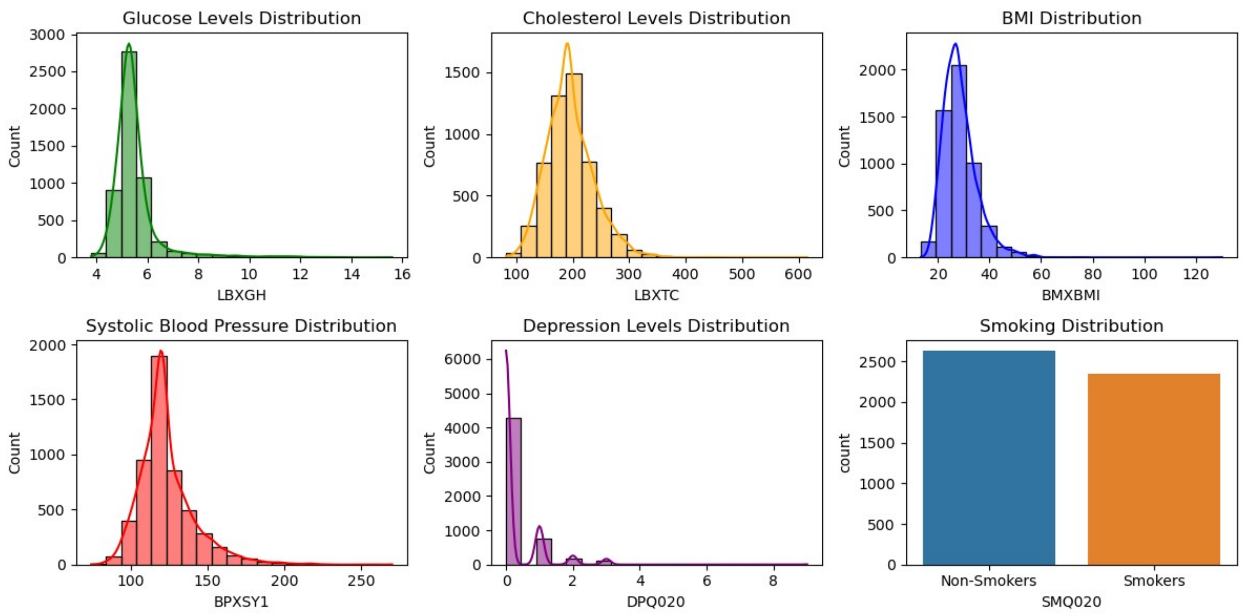
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from imblearn.over_sampling import SMOTE
from imblearn.ensemble import BalancedRandomForestClassifier
from sklearn.linear_model import LogisticRegression

import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Load the NHANES datasets for 2005-2006
glucose_data = pd.read_sas(r'GHB_D.XPT')          # Glycohemoglobin
(Glucose) Lab Data
cholesterol_data = pd.read_sas(r'TCHOL_D.XPT')    # Total Cholesterol
Lab Data
demo_data = pd.read_sas(r'BMX_D.XPT')            # Demographic Data
(includes BMI, Weight)
bp_data = pd.read_sas(r'BPX_D.XPT')               # Blood Pressure
Data
diet_data = pd.read_sas(r'DR1TOT_D.XPT')          # Dietary Data
(Total Nutrient Intakes - Day 1)
smoking_data = pd.read_sas(r'SMQ_D.XPT')          # Smoking
Questionnaire
mental_health_data = pd.read_sas(r'DPQ_D.XPT')    # Depression
Questionnaire
healthstatus_data = pd.read_sas(r'HSQ_D.XPT')     # Self-reported
Current Health Status

# Mapping data to for better understanding
healthstatus_data['HSD010'] =
healthstatus_data['HSD010'].replace({1:10, 2:8, 3:6, 4:2, 5:1, 6:1,
9:1})
smoking_data['SMQ020'] = smoking_data['SMQ020'].replace({1:1, 2:0,
7:0, 9:0})

# Merge datasets on 'SEQN'
merged_data = pd.merge(glucose_data, cholesterol_data, on='SEQN',
how='inner')
merged_data = pd.merge(merged_data, demo_data, on='SEQN', how='inner')
merged_data = pd.merge(merged_data, bp_data, on='SEQN', how='inner')
merged_data = pd.merge(merged_data, diet_data, on='SEQN', how='inner')
merged_data = pd.merge(merged_data, smoking_data, on='SEQN',
how='inner')
merged_data = pd.merge(merged_data, mental_health_data, on='SEQN',
```

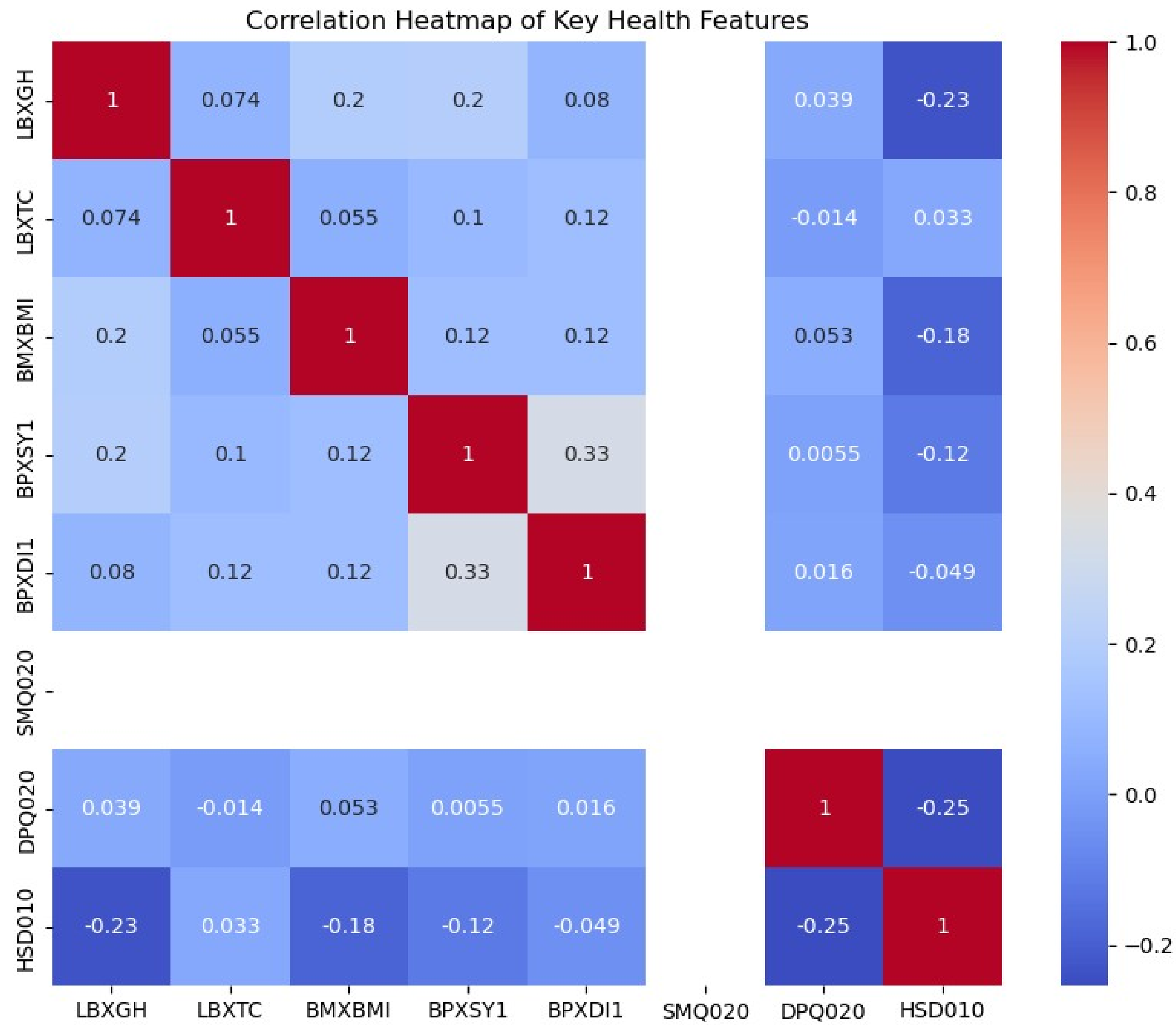
```
how='inner')
merged_data = pd.merge(merged_data, healthstatus_data, on='SEQN',
how='inner')

# Replacing missing values for cleaner analysis
features = ['LBXGH', 'LBXTC', 'BMXBMI', 'BPXSY1', 'BPXDI1', 'SMQ020',
'DPQ020']
final_data = merged_data[['LBXGH', 'LBXTC', 'BMXBMI', 'BPXSY1',
'BPXDI1', 'SMQ020', 'DPQ020', 'HSD010']]
final_data.fillna(0, inplace = True)

for feature in final_data.columns:
    median_value = np.median(final_data[feature][final_data[feature] !
= 0])  # Calculate median excluding zeros
    final_data[feature] = final_data[feature].replace(0, median_value)


smoking_data['SMQ020'] = smoking_data['SMQ020'].astype('category')

# Exploratory Data Analysis (EDA)
# Distribution of key health-related features
plt.figure(figsize=(12, 6))
plt.subplot(2, 3, 1)
sns.histplot(final_data['LBXGH'], bins=20, kde=True, color='green')
plt.title('Glucose Levels Distribution')

plt.subplot(2, 3, 2)
sns.histplot(final_data['LBXTC'], bins=20, kde=True, color='orange')
plt.title('Cholesterol Levels Distribution')

plt.subplot(2, 3, 3)
sns.histplot(final_data['BMXBMI'], bins=20, kde=True, color='blue')
plt.title('BMI Distribution')

plt.subplot(2, 3, 4)
sns.histplot(final_data['BPXSY1'], bins=20, kde=True, color='red')
plt.title('Systolic Blood Pressure Distribution')

plt.subplot(2, 3, 5)
sns.histplot(final_data['DPQ020'], bins=20, kde=True, color='purple')
plt.title('Depression Levels Distribution')

plt.subplot(2, 3, 6)
sns.countplot(x='SMQ020', data=smoking_data, hue = 'SMQ020', legend =
False)
plt.xticks([0, 1], ['Non-Smokers', 'Smokers'])
plt.title('Smoking Distribution')

plt.tight_layout()
plt.show()
```

```
# Correlation Heatmap for Key Health Features
plt.figure(figsize=(10, 8))
sns.heatmap(final_data[['LBXGH', 'LBXTC', 'BMXBMI', 'BPXSY1',
'BPXDI1', 'SMQ020', 'DPQ020', 'HSD010']].corr(), annot=True,
cmap='coolwarm')
plt.title('Correlation Heatmap of Key Health Features')
plt.show()

# Feature Engineering: Creating health indicators and target variables
# Target variable for diabetes (Glucose > 126 mg/dL)
final_data['diabetes'] = final_data['LBXGH'] > 126

# Cholesterol target (Total cholesterol > 200 mg/dL)
final_data['high_cholesterol'] = final_data['LBXTC'] > 200

# Hypertension target (Blood Pressure: Systolic > 130 or Diastolic >
80)
final_data['high_bp'] = (final_data['BPXSY1'] > 130) |
(final_data['BPXDI1'] > 80)

# Obesity indicator (BMI >= 30)
final_data['obesity'] = final_data['BMXBMI'] >= 30

# Smoking status (1 = Yes)
final_data['smoker'] = final_data['SMQ020'] == 1

# Depression level (PHQ-9 score, 0-27 scale)
final_data['depression'] = final_data['DPQ020']

# Health Score Calculation (Weighted combination of health factors,
scaled to 10)
weights = {
    'LBXGH': 0.2,  # Glucose
    'LBXTC': 0.15, # Cholesterol
    'BMXBMI': 0.15,# BMI
    'BPXSY1': 0.2, # Systolic Blood Pressure
    'BPXDI1': 0.1, # Diastolic Blood Pressure
    'SMQ020': 0.1, # Smoking
    'DPQ020': 0.1   # Depression
}

# Normalization function
def normalize(column):
    return (column - column.min()) / (column.max() - column.min())

# Normalize each feature and calculate the weighted score
final_data['health_score'] = (
    weights['LBXGH'] * (1 - normalize(final_data['LBXGH'])) +  # Lower
is better
```

```
    weights['LBXTC'] * (1 - normalize(final_data['LBXTC'])) +  # Lower
is better
    weights['BMXBMI'] * (1 - normalize(final_data['BMXBMI'])) + #
Lower is better
    weights['BPXSY1'] * (1 - normalize(final_data['BPXSY1'])) + #
Lower is better
    weights['BPXDI1'] * (1 - normalize(final_data['BPXDI1'])) + #
Lower is better
    weights['SMQ020'] * (1 - final_data['SMQ020']) +      # Higher is
better
    weights['DPQ020'] * (1 - normalize(final_data['DPQ020']))  # Lower
is better
)

# Scale health score to 0-10 range
final_data['health_score'] = final_data['health_score'] * 10

# Clip health score to ensure it's in the range [0, 10]
final_data['health_score'] = final_data['health_score'].clip(lower=0,
upper=10)

# Prepare data for machine learning
X = final_data[features]
y = final_data['HSD010']

# Scaling Data
scaler = MinMaxScaler()  # You can use StandardScaler() instead if
needed
X_scaled = scaler.fit_transform(X)
X_scaled = pd.DataFrame(X_scaled, columns=X.columns, index=X.index)

# Split the data for Machine Learning
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
test_size=0.3, random_state=42, stratify=y)

# Machine Learning Model (Random Forest for Classification)
model1 = RandomForestClassifier(random_state=42)
model1.fit(X_train, y_train)

# Machine Learning Model (Balanced Random Forest for Classification)
model2 = BalancedRandomForestClassifier(random_state=42)
model2.fit(X_train, y_train)

# Machine Learning Model (Logistic Regression for Classification)
model3 = LogisticRegression()
model3.fit(X_train, y_train)

# Predictions and evaluation
y_pred1 = model1.predict(X_test)
y_pred2 = model2.predict(X_test)
```

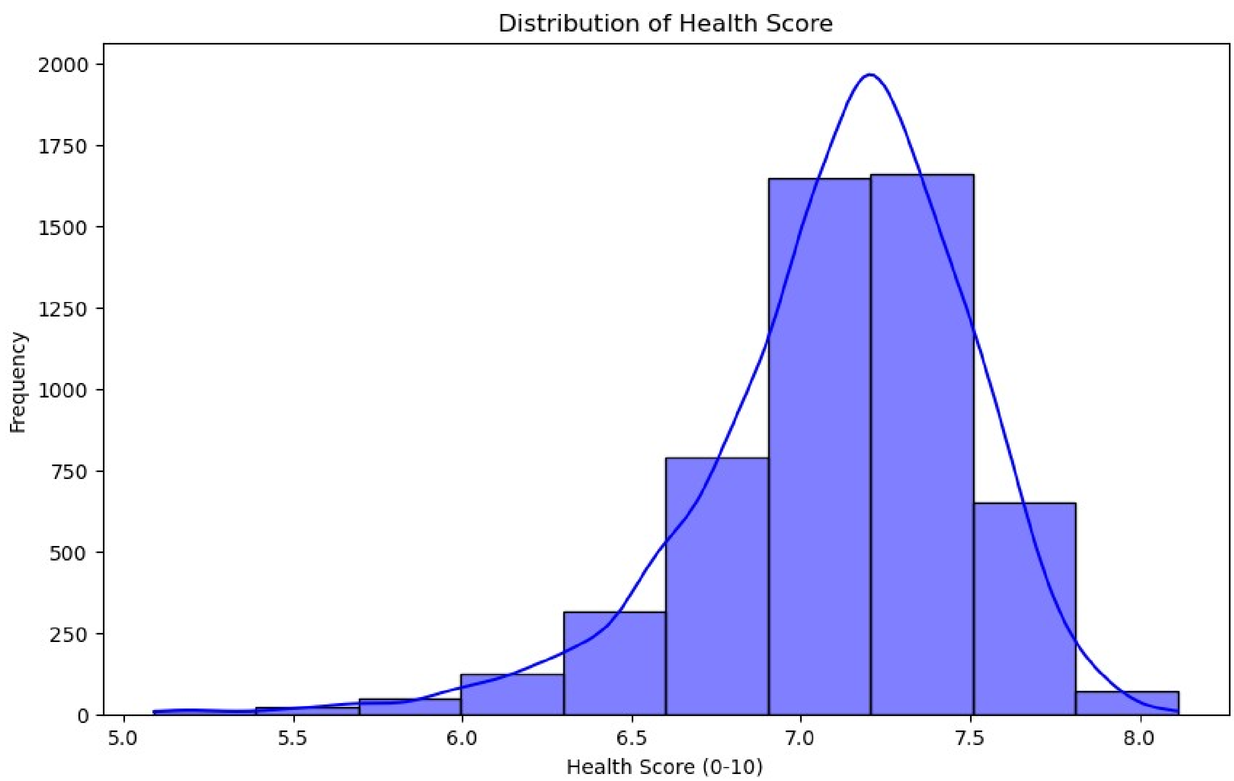
```
y_pred3 = model3.predict(X_test)

print('Classification Report for RandomForestClassifier: \n',
classification_report(y_test, y_pred1))
print('Classification Report for BalancedRandomForestClassifier: \n',
classification_report(y_test, y_pred2))
print('Classification Report for Logistic Regression: \n',
classification_report(y_test, y_pred3))

# Basic Visualization: Distribution of Health Score
plt.figure(figsize=(10, 6))
sns.histplot(final_data['health_score'], bins=10, kde=True,
color='blue')
plt.title('Distribution of Health Score')
plt.xlabel('Health Score (0-10)')
plt.ylabel('Frequency')
plt.show()
```






```
Classification Report for RandomForestClassifier:
              precision    recall  f1-score   support

         1.0       0.17      0.02      0.04        43
         2.0       0.36      0.19      0.25       263
         6.0       0.46      0.70      0.56       713
         8.0       0.33      0.27      0.30       434
        10.0       0.11      0.01      0.02       148

    accuracy                           0.42      1601
   macro avg       0.29      0.24      0.23      1601
weighted avg       0.37      0.42      0.37      1601

Classification Report for BalancedRandomForestClassifier:
              precision    recall  f1-score   support

         1.0       0.06      0.42      0.11        43
```

```
         2.0       0.25      0.27      0.26       263
         6.0       0.51      0.17      0.26       713
         8.0       0.36      0.26      0.30       434
        10.0       0.13      0.41      0.20       148

    accuracy                           0.24      1601
   macro avg       0.26      0.31      0.22      1601
weighted avg       0.38      0.24      0.26      1601

Classification Report for Logistic Regression:
               precision    recall  f1-score   support

         1.0       1.00      0.02      0.05        43
         2.0       0.36      0.08      0.13       263
         6.0       0.45      0.97      0.62       713
         8.0       0.11      0.00      0.00       434
        10.0       0.00      0.00      0.00       148

    accuracy                           0.45      1601
   macro avg       0.38      0.22      0.16      1601
weighted avg       0.32      0.45      0.30      1601
```



For the HealthSage app, additional features from the NHANES dataset can be extremely useful in creating a more comprehensive health assessment model. NHANES collects a wide range of health-related data, allowing you to incorporate diverse features related to lifestyle, environment, mental health, and other biological markers.

Here's a breakdown of **additional features** that can enhance the predictive capability of the **HealthSage** app:

15.1 Biochemical Data

- **Blood Glucose Levels (Glycohemoglobin, Fasting Glucose)**: Already used, but further breakdown into **HbA1c** (glycosylated hemoglobin) can help with long-term diabetes risk prediction.
- **Cholesterol Types**: Not just LDL, but also **HDL** (High-Density Lipoprotein), **Total Cholesterol**, and **Triglycerides** for a more detailed heart disease risk profile.
- **C-Reactive Protein (CRP)**: An indicator of inflammation, linked to cardiovascular and chronic diseases.
- **Liver Enzymes (ALT, AST)**: Can be used to evaluate liver function, which could be integrated into overall health scoring.
- **Blood Urea Nitrogen (BUN) and Creatinine**: Indicators of kidney function, which can affect general health.

15.2 Dietary Data

- **Daily Nutrient Intake**: NHANES contains detailed **24-hour dietary recall** data, which tracks the daily intake of:
  - **Macronutrients**: Protein, Carbohydrates, Fats
  - **Micronutrients**: Vitamins (A, C, D), Minerals (Iron, Calcium)
- **Food Frequency Questionnaire**: This assesses the consumption of various food groups over time, which can influence disease risks like heart disease and diabetes.

14.3 Physical Activity & Fitness

- **Activity Monitors (Steps, Heart Rate)**: Already incorporated from wearable devices, but additional metrics like **sedentary time**, **vigorous activity**, and **exercise frequency** can give a more holistic view of physical fitness.
- **Muscle Strength Tests**: NHANES includes **grip strength** as an indicator of overall muscle strength and functional health.

14.4 Body Measurements

- **Waist Circumference**: A predictor of metabolic syndrome and cardiovascular risk.
- **Hip Circumference**: Used alongside waist circumference to calculate the **Waist-to-Hip Ratio**, a strong indicator of obesity-related health risks.
- **Body Fat Percentage**: Available in some years, offering a more precise measure than BMI for obesity-related health risks.

14.5 Sleep Data

- **Sleep Questionnaire**: Tracks sleep duration, quality, and disorders (e.g., **sleep apnea**), which is essential for predicting long-term health outcomes like cardiovascular disease and obesity.

14.6 Smoking & Alcohol Use

- **Smoking Status**: Self-reported data on current or former smoking, which has a direct impact on lung health, heart disease, and cancer risk.
- **Alcohol Consumption**: Data on frequency and quantity of alcohol intake, which is crucial for liver health, mental well-being, and overall disease risk.

14.7 Mental Health Data

- **Depression Symptoms**: NHANES provides mental health data through standardized questionnaires like **PHQ-9** (Patient Health Questionnaire), which assesses depression severity.
- **Stress Levels**: Some NHANES cycles contain data on **stress**, **anxiety**, and **emotional well-being**.

14.8 Environmental Exposure

- **Exposure to Pollutants**: NHANES tracks levels of certain **environmental chemicals** and **heavy metals** in the blood (e.g., **lead**, **mercury**, **pesticides**), which can influence long-term health risks.
- **Secondhand Smoke Exposure**: This can be important for assessing respiratory and cardiovascular risks.

14.9 Chronic Conditions & Medications

- **Self-Reported Health Conditions**: Data on whether an individual has been diagnosed with conditions like **diabetes**, **heart disease**, **hypertension**, **asthma**, and **cancer**.

- **Prescription Medication Use**: Tracks the use of medications, which can provide insights into existing health conditions and management strategies.

14.10 Cognitive Function

- **Cognitive Tests**: NHANES includes tests related to **memory**, **processing speed**, and **problem-solving**, which can indicate cognitive decline and overall brain health.
- **Hearing Tests**: Indicators of hearing loss, which can be linked to cognitive decline.

14.11 Bone Health

- **Osteoporosis Indicators**: Bone density measures, such as **dual-energy X-ray absorptiometry (DXA)** scans, can be used to predict osteoporosis and fracture risks.

15.0 Conclusion:

The proposed AI-powered health prediction app addresses a significant need in India's healthcare ecosystem by transforming raw lab data into predictive insights and actionable health recommendations. This app, leveraging AI and big data, will be an essential tool in preventive healthcare, allowing users to understand and mitigate future health risks proactively. It stands out from competitors like Lab Informer and Docus by offering a more holistic approach, integrating lifestyle data, genetic information, and personalized preventive care, providing a comprehensive, AI-powered health assessment for the Indian market.

References:

1. **Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011**. Retrieved from:
   - https://www.meity.gov.in/writereaddata/files/it_act_2000/IT_Sensitive_Personal_Data_Rules_2011.pdf
2. **Personal Data Protection Bill, 2019**. Retrieved from:
   - https://prsindia.org/billtrack/personal-data-protection-bill-2019
3. **Medical Device Rules, 2017**. Retrieved from:
   - https://cdsco.gov.in/opencms/opencms/system/modules/CDSCO.WEB/elements/download_file_division.jsp?num_id=MzA2OA==
4. **Indian Patents Act, 1970**. Retrieved from:
   - https://legislative.gov.in/sites/default/files/A1970-39.pdf
5. **Lab Informer**. Retrieved from:
   - https://www.labinformer.com/
6. **Docus AI**. Retrieved from:
   - https://docus.ai/lab-test-interpretation

7. **HealthifyMe - Nutrition & Diet App**. Retrieved from:
    - https://www.healthifyme.com/
8. **Medical Council of India (MCI) Regulations**. Retrieved from:
    - https://www.nmc.org.in/acts-and-amendments/mci-regulations/

Concept Generation:

1. **OpenAI. (2024).** *ChatGPT* (Version 4) [AI model].
   Retrieved from https://www.openai.com/chatgpt
2. **Wearable Health Devices and AI in Healthcare**:
    - Masud, M., Gaba, G. S., et al. "Wearable Sensors for Remote Health Monitoring: Current Status and Future Challenges." *Sensors*, 2022.
        - https://www.mdpi.com/1424-8220/22/4/1169
3. **AI in Personalized Healthcare**:
    - Jha, S., Topol, E. J. "Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists." *JAMA*, 2016.
        - https://jamanetwork.com/journals/jama/fullarticle/2556071
4. **AI for Preventive Healthcare**:
    - Yoo, I., Alafaireet, P., et al. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature." *Journal of Medical Systems*, 2012.
        - https://link.springer.com/article/10.1007/s10916-011-9710-5
5. **Lifestyle Data and AI Health Predictions**:
    - Fagherazzi, G., Ravaud, P. "Digital Health Technologies and Diabetes Management: A User's Guide." *The Lancet Diabetes & Endocrinology*, 2019.
        - https://www.thelancet.com/journals/landia/article/PIIS2213-8587(19)30026-6/fulltext

Market Need:

1. **Market Size and Growth of Preventive Healthcare**:
    - Grand View Research. "Preventive Healthcare Market Size, Share & Trends Analysis Report by Type (Regular Health Check-ups, Vaccination), by Region (North America, APAC), and Segment Forecasts, 2023-2030."
        - https://www.grandviewresearch.com/industry-analysis/preventive-healthcare-market
2. **Indian Healthcare and Disease Statistics**:
    - World Health Organization (WHO). "Noncommunicable Diseases Country Profiles 2022 – India."
        - https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
3. **Consumer Trends in Healthcare**:

- Accenture. "Accenture 2021 Health & Life Sciences Experience Survey: Responding to Consumer Health Needs."
    - https://www.accenture.com/us-en/insights/health/consumer-health-needs-survey

4. **Rise in Chronic Diseases in India**:
    - Indian Council of Medical Research (ICMR). "ICMR's Longitudinal Ageing Study in India Report (LASI) 2020: Chronic Disease Trends in India."
        - https://main.icmr.nic.in/sites/default/files/whats_new/ICMR_LASI_Report_2020.pdf