HealthPredict Pro

Dhara Khamar
Mayank Pujara
Vaishnavi
Chardra Shekhar
Swarnabha

3rd March, 2024

Abstract

HealthPredict Pro is a state-of-the-art predictive analytics application for disease detection and personalized healthcare. Leveraging advanced machine learning and digital health technologies, it offers accurate predictions for diseases like pneumonia, mental health disorders, and diabetes by analyzing diverse personal health data sources. The application aims to monetize through subscription services, licensing, and strategic partnerships while ensuring regulatory compliance and user trust. With a user-friendly interface and commitment to continuous innovation, HealthPredict Pro empowers individuals to improve their health outcomes effectively.

1. Introduction

a. Context

The development of an application capable of detecting and predicting diseases based on machine learning models is a significant advancement in the field of healthcare and technology. With the proliferation of digital health solutions and the increasing availability of personal health data, there is a growing opportunity to leverage machine learning algorithms to improve disease detection and prediction.
The context for developing an application capable of detecting and predicting diseases based on machine learning models arises from several converging factors in the healthcare and technology domains:

1. Advancements in Machine Learning: Rapid progress in machine learning, especially deep learning algorithms, enables accurate pattern recognition and analysis, ideal for disease detection and prediction.
2. Digital Health Revolution: The proliferation of digital health technologies provides vast amounts of health data from sources like wearables and EHRs, offering opportunities for data-driven healthcare solutions.
3. Rising Disease Burden: Diseases such as pneumonia, mental health disorders, and diabetes pose significant global health challenges, necessitating early detection and intervention.
4. Personalized Medicine Paradigm: Machine learning allows for the development of predictive models tailored to individual genetics, lifestyle, and environmental factors, supporting personalized healthcare approaches.
5. Healthcare Access and Equity: Digital health solutions, including predictive analytics applications, can improve healthcare access and equity by providing remote monitoring and decision support tools.
6. Preventive Healthcare Focus: Emphasis on preventive healthcare strategies underscores the importance of early detection and intervention, facilitated by predictive analytics applications to identify high-risk individuals.
7. Regulatory Landscape: Developers must navigate regulatory requirements concerning data privacy, security, and ethical considerations to ensure compliance with relevant laws and standards in healthcare technology development.

b. Purpose & scope

Purpose:
The purpose of this application is to provide individuals with a convenient and accessible tool for early detection and prediction of certain diseases, namely pneumonia, mental health disorders, and diabetes. By analyzing personal health data provided by patients, the application aims to identify potential risks and offer timely interventions or recommendations for further medical assessment.

Scope:
The application's scope encompasses the development of machine learning models tailored to each specific disease - pneumonia, mental health disorders, and diabetes. These models will be trained on relevant datasets containing diverse patient data, including demographic information, medical history, lifestyle factors, and diagnostic tests. The application will feature a user-friendly interface where individuals can input their personal data securely and receive predictions or risk assessments for the targeted diseases.

c. Objective

Develop a predictive analytics application for disease detection and personalized healthcare, leveraging machine learning and digital health tech. Aim to monetize through subscription services, licensing, or partnerships while ensuring regulatory compliance. Drive user engagement, market penetration, and continuous innovation to support sustainable growth for healthcare support startups.

2. Market/ Customer / Business Need Assessment

- **Market Assessment:**
  - Size and Growth: Each sector represents a significant market: Diabetes management ($20 billion globally), Mental Health Solutions (significant market in the USA with approximately 51.5 million affected adults annually), and Pneumonia diagnostics (significant demand due to morbidity and mortality impact).
  - Trends: Trends include the integration of technology, such as AI-driven analytics and teletherapy, increasing awareness and destigmatization, and a focus on personalized and accurate diagnostics and treatments.
  - Competition: Each sector faces competition from established players and startups offering innovative solutions.
  - Regulations: Compliance with FDA, HIPAA, and other regulatory bodies is crucial for market approval and deployment of solutions.

- **Business Needs Assessment:**
  - Data Access and Acquisition: Secure high-quality healthcare data through partnerships with providers and research institutions.
  - Expertise: Build multidisciplinary teams with expertise in machine learning, healthcare, regulatory compliance, and technology development.
  - Partnerships and Collaborations: Collaborate with healthcare providers, insurers, employers, and government agencies to pilot, validate, and deploy solutions.
  - Revenue Models: Explore subscription-based services, licensing agreements, direct-to-consumer sales, and reimbursement strategies.
  - Technology Infrastructure: Invest in secure platforms, data analytics, and electronic health record (EHR) systems while ensuring compliance with regulations like HIPAA.
  - Clinical Validation and Quality: Conduct clinical trials and validation studies to demonstrate efficacy, safety, and clinical outcomes.
  - Market Adoption and Commercialization: Develop robust commercialization strategies, including pricing, sales, marketing efforts, and engagement with key stakeholders.

- **Customer Characterization:**
  - Patients/Clients: Individuals affected by diabetes, mental health conditions, or pneumonia, seeking accurate diagnostics, personalized treatments, and support.
  - Caregivers and Family Members: Responsible for managing and supporting patients, seeking tools and resources for effective caregiving.
  - Healthcare Providers: Including doctors, nurses, therapists, radiologists, and other specialists, seeking innovative solutions to improve diagnosis, treatment, and patient outcomes.
  - Employers and Insurers: Interested in promoting employee well-being, reducing healthcare costs, and implementing cost-effective management and prevention strategies.
  - Government and Public Health Agencies: Concerned with public health surveillance, outbreak detection, prevention strategies, and regulation of healthcare solutions.

3. External Search & Benchmarking

Statistical Data:

https://www.cdc.gov/diabetes/data/index.html
https://www.cdc.gov/nchs/nhanes/index.htm
The National Health and Nutrition Examination Survey (NHANES) is a national survey that monitors the health and nutritional status of adults and children across the United States.
http://data.ctdata.org/dataset/mental-health
https://data.unicef.org/topic/child-health/pneumonia/

Relevant Papers:

- Srividya M., Mohanavalli S., Bhalaji N. Behavioral Modeling for Mental Health using Machine Learning Algorithms. J. Med. Syst. 2018;42:88. doi: 10.1007/s10916-018-0934-
- Milne-Ives M., Lam C., De Cock C., Van Velthoven M.H., Meinert E. Mobile Apps for Health Behavior Change in Physical Activity, Diet, Drug and Alcohol Use, and Mental Health: Systematic Review. JMIR mHealth uHealth. 2020;8:e17046. doi: 10.2196/17046
- Liu C., McCabe M., Dawson A., Cyrzon C., Shankar S., Gerges N., Kellett-Renzella S., Chye Y., Cornish K. Identifying Predictors of University Students' Wellbeing during the COVID-19 Pandemic—A Data-Driven Approach. Int. J. Environ. Res. Public Health. 2021;18:6730. doi: 10.3390/ijerph18136730
- Naser M. Mapping functions: A physics-guided, data-driven and algorithm-agnostic machine learning approach to discover causal and descriptive expressions of engineering phenomena. Measurement. 2021;185:110098. doi: 10.1016/j.measurement.2021.110098
- Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord.* 19 (1), 101–109. doi:10.1186/s12902-019-0436-6
- Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. *SN Appl. Sci.* 1 (12), 1667–1712. doi:10.1007/s42452-019-1759-7
- Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., and Kumar, M. (2021). eDiaPredict: an ensemble-based framework for diabetes prediction. *ACM Trans. Multimedia Comput. Commun. Appl.* 17 (2), 1–26. doi:10.1145/3415155
- Stephen O., Sain M., Maduh U. &Jeong D. An efficient deep learning approach to pneumonia classification in healthcare. Journal Of Healthcare Engineering. 2019 (2019) https://doi.org/10.1155/2019/4180949 PMID: 31049186
- Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv: 2003.10849. 2020.
- Vaishya R, Javaid M, Khan IH, Haleem A. Artifcial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clin Res Rev. 2020.

4. Business Model

The business model will be based on a sophisticated **subscription-based platform** as its pricing structure for individuals and **variable pricing structure** for corporate/institutions. The customers will be given different options to select from:

➢ **Individual Plans:**

Free Tier:

- Limited functionality: Offer basic features in one health area (e.g., basic pneumonia risk assessment or mood tracking for mental health).
- Include educational resources and introductory content.
- Serves as a lead generation tool to convert users to paid plans.

Limited Plan (Monthly: ₹99, Annual: ₹899):

- Access to all features in one health area (full pneumonia detection with basic severity analysis/ comprehensive mental health assessment/ diabetes prediction).
- Limited consultations (1 free consultation per month) or discounted consultation rates.

- Ideal for users with a specific health concern.

Premium Plan (Monthly: ₹199, Annual: ₹1499):

- Full access to all features across all health areas (pneumonia detection, mental health assessment, diabetes prediction).
- Maximum 8 consultations per month (or significantly discounted rates) with mental health professionals.
- Personalized recommendations and progress tracking across health areas.

➢ **Corporate/Institutional Plans (Variable Pricing):**

Offer customizable packages based on the size of the organization and specific needs. Some of the possible program considerations are:

- Employee Wellness Programs: Discounted group subscriptions for mental health assessments and personalized recommendations.
- Student Support Services: Provide access to mental health resources and consultations for students at educational institutions.
- Hospital Partnerships: Integrate our platform with hospital systems for early diagnosis and improved patient outcomes.

5. Concept Generation & Concept Development

We have proposed a unified healthcare platform leveraging machine learning for three key functionalities:

a. Pneumonia Detection and Analysis:

Advanced Pneumonia Typing and Analysis:

- Differentiate between bacterial, viral, and other pneumonia types using advanced X-ray features.
- Predict complications like pleural effusion or lung abscesses, aiding in treatment decisions.
- Analyse disease severity and estimate the risk of hospitalization or need for intensive care.

Personalized Risk Assessment and Early Warning:

- Integrate chest X-rays with patient data for personalized risk assessments.
- Predict individual pneumonia risks, enabling preventive measures.
- Implement real-time monitoring for early detection of potential pneumonia onset.

Integrated Clinical Decision Support System:

- Seamlessly integrate AI tool within radiology workflows and hospital information systems.
- Provide clinicians with AI-powered insights and recommendations.
- Generate personalized treatment plans based on AI predictions and patient-specific factors.

Public Health and Surveillance Tool:

- Develop a system for automated screening of chest X-rays in large populations.
- Contribute to epidemic surveillance by identifying clusters of pneumonia cases in real-time.
- Facilitate resource allocation and public health interventions with early warnings.

b. Mental Health Disorders:

Comprehensive Mental Health Assessment:

- Utilize advanced machine learning algorithms for a comprehensive assessment.
- Analyse patient self-reports, behavioral patterns, and clinical history.

Early Detection and Intervention:

- Use predictive analytics to identify potential mental health concerns early.
- Provide early warnings and risk assessments for proactive intervention.

Personalized Treatment Plans:

- Assist clinicians in developing personalized treatment plans based on AI-driven insights.
- Consider individual characteristics, treatment history, and response to interventions.
- Enable patients to track mental health progress and provide clinicians with valuable data.

c. **Diabetes Prediction:**

Early Diabetes Risk Assessment:

- Employ advanced machine learning models to assess the risk of diabetes.
- Consider factors like genetic predisposition, lifestyle, and medical history.

Preventive Measures and Lifestyle Recommendations*:*

- Offer personalized recommendations for preventive measures and lifestyle modifications.
- Empower individuals to make informed choices and reduce the risk of diabetes.

Clinician Decision Support:

- Serve as a decision support tool for healthcare professionals.
- Provide predictive insights to tailor interventions and treatment plans.

Our healthcare solution basically focuses on advanced capabilities, personalization, integration, and real-time monitoring across pneumonia detection, mental health assessment, and diabetes prediction, ensuring a comprehensive and proactive healthcare solution.

Project Components:

1. Deep Learning Frameworks:
- TensorFlow: Core framework for building and training deep learning models.
- PyTorch: Dynamic computation graph for model flexibility, particularly in mental health assessment.

2. Model Architectures:
- VGG-16: Effective for pneumonia detection with a deep structure and small convolutional filters.
- Transfer Learning Models: Pre-trained models like ResNet or MobileNet for fine-tuning in diabetes prediction.
- Interpretable Algorithms: Decision trees, rule-based models, or explainable neural networks for transparent mental health assessment.

3. Data Processing and Augmentation:
- OpenCV: Used for image preprocessing tasks, including resizing, normalization, and noise removal.
- Scikit-image: Utilized for advanced image processing tasks to enhance feature extraction.

4. Data Management:
- Pandas: Efficient data manipulation and organization, especially for handling patient data.
- SQL Databases: Structured storage of patient information, ensuring quick access for machine learning modeling.

5. Collaboration and Version Control:
- Git and GitHub: Version control and collaborative development, facilitating integration of contributions from team members.

6. Model Evaluation and Interpretability:
- Scikit-learn: Employed for model evaluation tasks, metrics calculation, and cross-validation.
- LIME (Local Interpretable Model-agnostic Explanations): Enhances interpretability, particularly in mental health assessment.

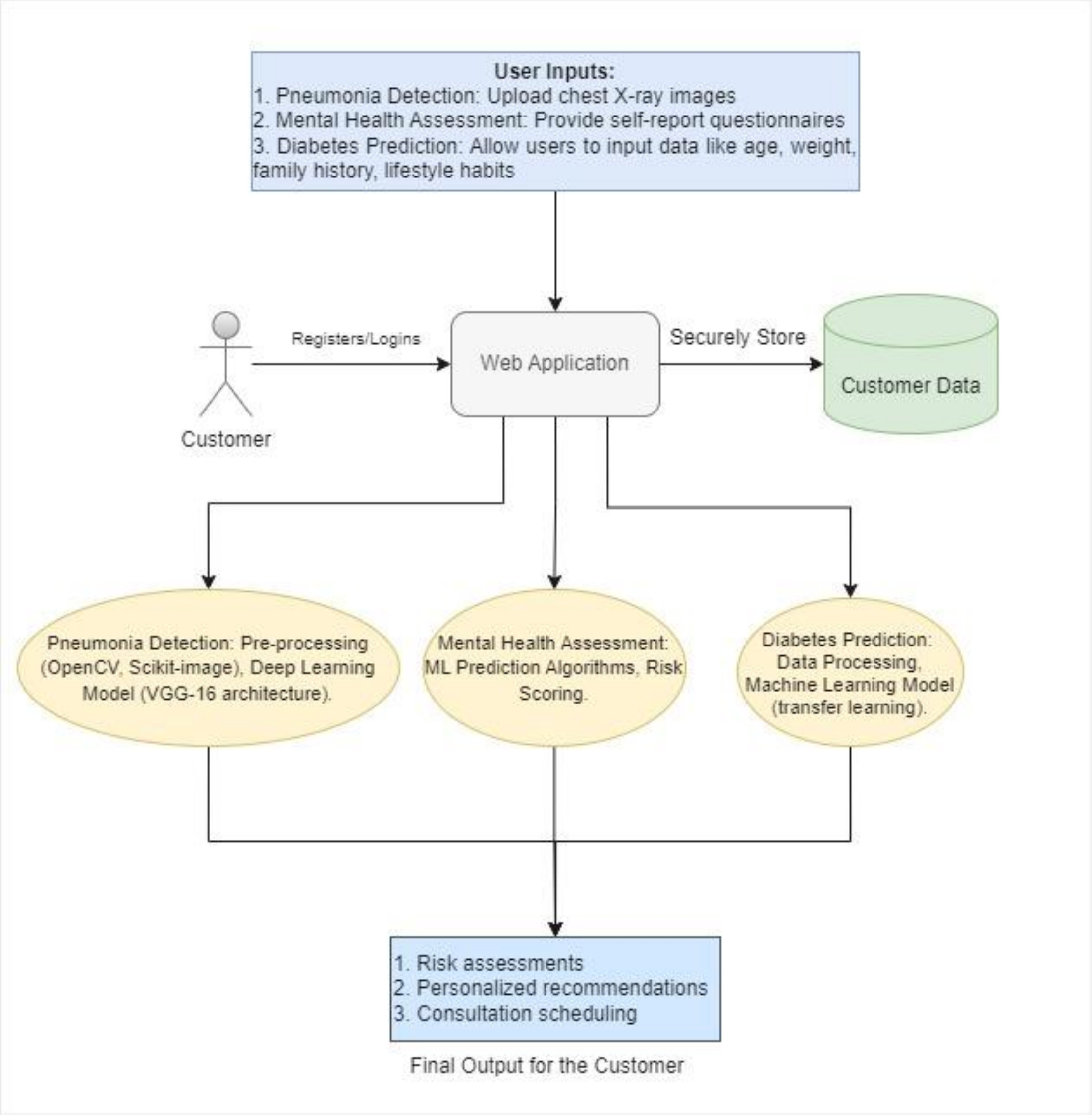
6. Final Product Prototype (abstract) with Schematic Diagram

Implementing the final prototype for our healthcare solution primarily involves designing and developing an intuitive user interface (UI) and a robust backend that seamlessly integrates the three key functionalities:

1. Pneumonia Detection and Analysis
2. Mental Health Disorders prediction
3. Diabetes Prediction

User Interface (UI) and User Experience (UX):

- Unified Platform: Design a user-friendly platform accessible through a web app
- Clear Navigation: Separate sections for Pneumonia Detection, Mental Health Assessment, and Diabetes Prediction.


User Inputs:
1. Pneumonia Detection: Upload chest X-ray images
2. Mental Health Assessment: Provide self-report questionnaires
3. Diabetes Prediction: Allow users to input data like age, weight, family history, lifestyle habits
Registers/Logins
Customer
Web Application
Securely Store
Customer Data
Pneumonia Detection: Pre-processing (OpenCV, Scikit-image), Deep Learning Model (VGG-16 architecture).
Mental Health Assessment: ML Prediction Algorithms, Risk Scoring.
Diabetes Prediction: Data Processing, Machine Learning Model (transfer learning).
1. Risk assessments
2. Personalized recommendations
3. Consultation scheduling
Final Output for the Customer

Software Functionalities:

a. Pneumonia Detection:

- User uploads a chest X-ray image. The system pre-processes the image using OpenCV and Scikit-image for noise removal and feature extraction.
- The deep learning model analyses the image.
- The system displays the results:
  - Probability of pneumonia (bacterial, viral, or other).
  - Likelihood of complications.
  - Disease severity and hospitalization/ICU risk.

b. Mental Health Assessment:

- User completes self-report questionnaires and logs moods/symptoms.
- The system analyses the data using machine learning algorithms tailored for mental health assessment, potentially including interpretable models for better understanding.
- The system provides:
    - Risk assessment for various mental health conditions.
    - Early warning signs for potential concerns.
    - Personalized recommendations for self-care or seeking professional help.
- Integrate with a scheduling system for online consultations with licensed therapists (included with Paid Plans).

c. Diabetes Prediction:

- User enters data on a dedicated section.
- The system utilizes a machine learning model (potentially using transfer learning) to assess diabetes risk.
- The system displays:
    - Individualized risk score for developing diabetes.
    - Personalized recommendations for preventive measures (diet, exercise).
    - Educational resources on diabetes prevention and management.

Overall, the software aims to provide users with actionable insights and personalized recommendations to improve their health outcomes. Further refinement of the business model and continuous development based on user feedback will be crucial for establishing this solution as a valuable tool in the healthcare landscape.

7. Code Implementation/Validation on Small Scale

a. Pneumonia Detection

- **Code Link:** https://github.com/CHANDRASHEKHAR2898/PNEUMONIA-DETECTION/blob/main/Pneumonia_Classification.ipynb

b. Global Trends in Mental Health Disorder

- **Code Link:** https://github.com/CHANDRASHEKHAR2898/PNEUMONIA-DETECTION/blob/main/Pneumonia_Classification.ipynb
- **Dataset Used**:

  The dataset which is used in this project is the Global Trends in Mental Health Disorder on Kaggle. You can access the dataset using: Dataset Link

  This dataset contains informative data from countries across the globe about the prevalence of mental health disorders including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression and alcohol use disorders. Each row of the table contains information about a certain country or region for a certain year.

  The following columns are provided:

  - Entity (the country or region name)
  - Code (the code for the country or region)
  - Year (the year the data was collected)
  - Schizophrenia (% - percentage of people with schizophrenia)
  - Bipolar Disorder (% - percentage of people with bipolar disorder)
  - Eating Disorders (%) - percentage of individuals with disordered eating patterns
  - Anxiety Disorders (%) - percentage of individuals with anxiety
  - Drug Use Disorders (%) - percentage figures for those struggling with substance abuse
  - Depression (%) – percentages relating to those struggling with depressive illness
  - Alcohol Use Disorders (%) – percentages relating to those battling alcoholism

- **Data Preprocessing:**

  The dataset is loaded using pandas, and basic information like the column names, data types, and shape of the dataset is checked. The data preprocessing steps undertaken in the code are:

  a. Handling missing values:
  - Identified missing values in specific columns using pd.isnull(dataset).sum().
  - Dropped rows with missing values in key columns (Code, mental disorder percentage columns) using dropna(subset=missing_values_columns, inplace=True).

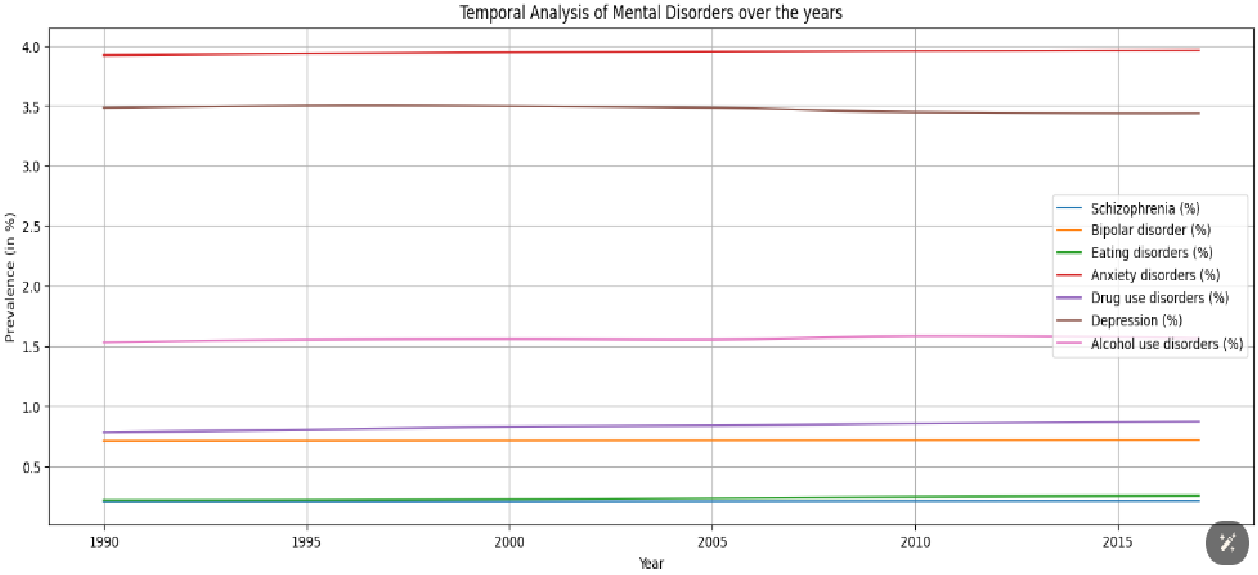
  b. Handling data types:
  - Converted numeric columns to numeric data type using pd.to_numeric(errors='coerce') to handle potential non-numeric entries.

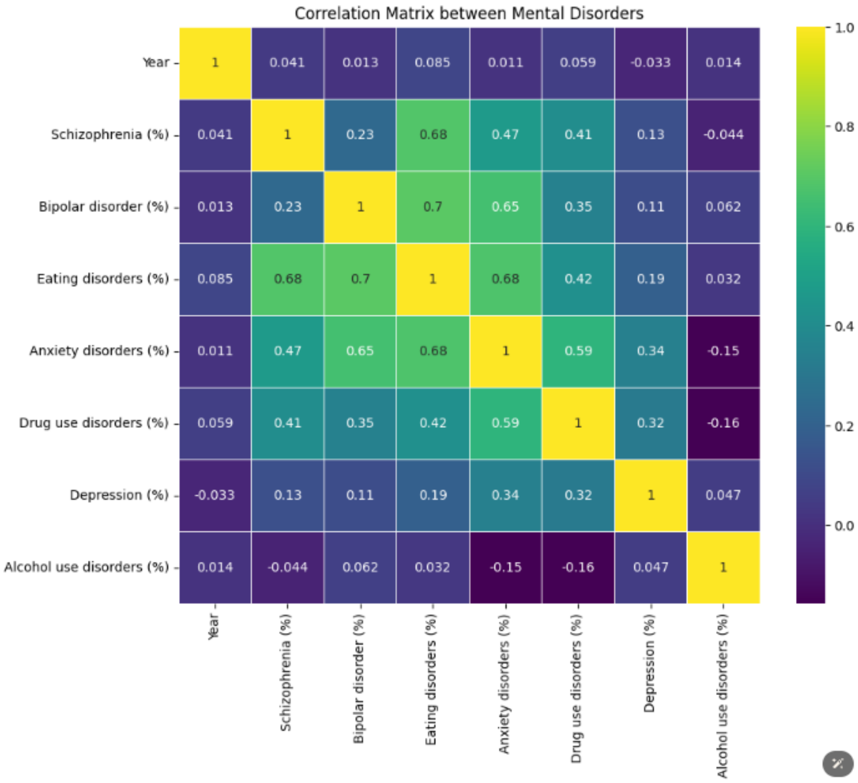
  c. Resetting index:
  - Reset the index after dropping rows to maintain proper indexing using reset_index(drop=True, inplace=True).

- **Exploratory Data Analysis (EDA):**

  Key findings from the EDA include:

  - **Temporal analysis:** Yearly mean trends of various mental disorders were visualized using a line chart, potentially revealing temporal patterns or changes.
  - **Global prevalence:** World maps were generated using *geopandas* and *folium* to depict the global prevalence of each disorder by country, offering insights into geographical variations.
  - **Correlation matrix:** A correlation matrix was created using *sns.heatmap* to identify relationships between different mental disorders.





This revealed:

a. Moderate positive correlations between some disorders (bipolar disorder and Eating Disorders, Eating Disorders and Anxiety Disorders).

b. A weak positive correlation between Schizophrenia and other disorders.

c. A weak negative correlation between Alcohol Use Disorders and most other disorders.

➢ Model Building:

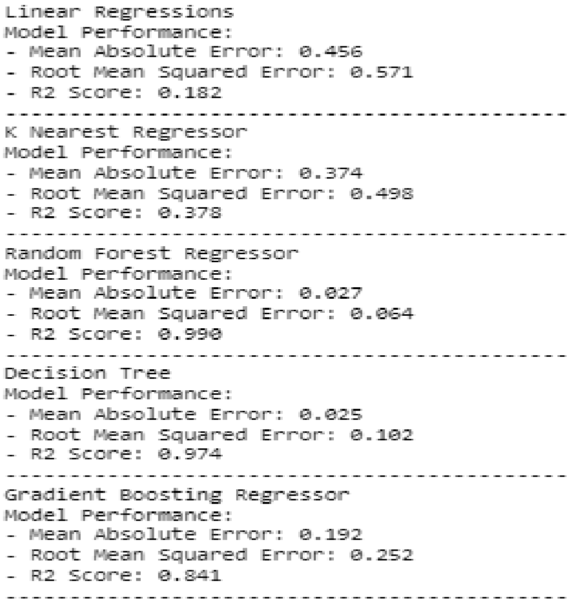
The following model building steps were implemented:

- **Feature selection:** Relevant features for predicting depression prevalence were chosen, including other mental disorder percentages and year.
- **Data splitting:** The data was split into training and testing sets using *train_test_split* to evaluate model performance on unseen data.
- **Model definition:** Five different regression models were defined:
  1. Linear Regression
  2. K Neighbours Regressor

   3. Random Forest Regressor
   4. Decision Tree Regressor
   5. Gradient Boosting Regressor

- **Model fitting:** Each model was fit to the training data using *model.fit(X_train, y_train)*.
- **Prediction:** Predictions were made on the testing data using *model.predict(X_test)*.
- **Evaluation:** The performance of each model was evaluated using the following metrics:
   a) Mean Absolute Error (MAE)
   b) Root Mean Squared Error (RMSE)
   c) R2 Score

➢ Evaluation:



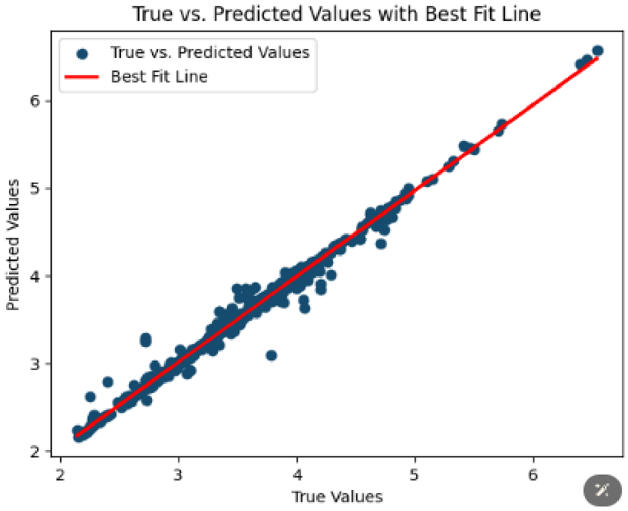
```
Linear Regressions
Model Performance:
- Mean Absolute Error: 0.456
- Root Mean Squared Error: 0.571
- R2 Score: 0.182
--------------------------------------------
K Nearest Regressor
Model Performance:
- Mean Absolute Error: 0.374
- Root Mean Squared Error: 0.498
- R2 Score: 0.378
--------------------------------------------
Random Forest Regressor
Model Performance:
- Mean Absolute Error: 0.027
- Root Mean Squared Error: 0.064
- R2 Score: 0.990
--------------------------------------------
Decision Tree
Model Performance:
- Mean Absolute Error: 0.025
- Root Mean Squared Error: 0.102
- R2 Score: 0.974
--------------------------------------------
Gradient Boosting Regressor
Model Performance:
- Mean Absolute Error: 0.192
- Root Mean Squared Error: 0.252
- R2 Score: 0.841
--------------------------------------------
```

To determine the best model the following factors must be evaluated:

- **Mean Absolute Error (MAE):** Lower MAE values indicate better accuracy.
- **Root Mean Squared Error (RMSE):** Lower RMSE values indicate better accuracy, similar to MAE.
- **R2 Score:** A higher R2 score indicates a better fit of the model to the data.

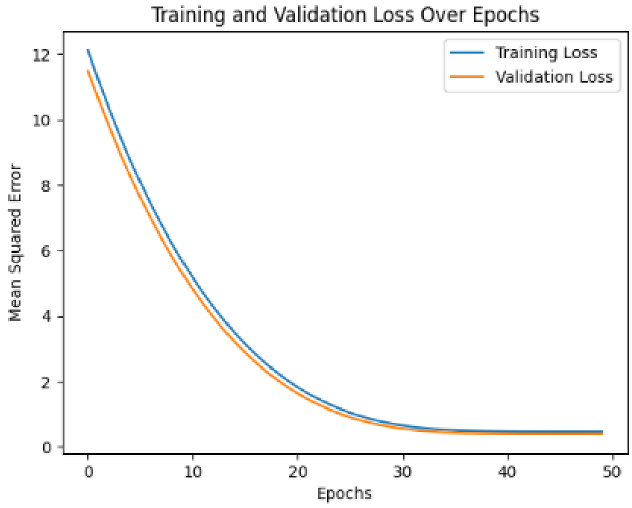
Considering all three metrics (MAE, RMSE, R2 Score), **the Random Forest Regressor** consistently outperforms other models in terms of accuracy, precision, and overall fit to the data. It has the lowest MAE and RMSE values, indicating better accuracy, and the highest R2 score, indicating a better fit to the data.

Therefore, based on the provided metrics, **the Random Forest Regressor** appears to be the best model among the ones evaluated.



Neural Network Implementation:

- In addition to traditional machine learning approaches, a neural network model was implemented for predicting depression prevalence based on various mental health indicators.
- The preprocessing involved standard scaling of the features to enhance model performance. The neural network architecture consisted of multiple dense layers, culminating in a linear activation function for regression.
- The model was compiled using the Adam optimizer and mean squared error loss function.
- Training over 50 epochs, the neural network demonstrated its ability to learn complex relationships within the data.
- Evaluation metrics such as mean squared error were employed to assess its predictive performance on the test set.



- The neural network implementation offers an alternative avenue for exploring intricate nonlinear patterns in mental health data, presenting an opportunity for further refinement and exploration in future studies.

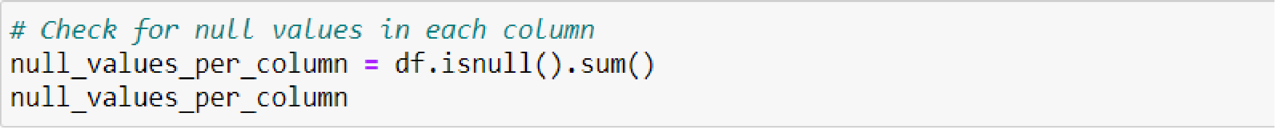
- **Conclusion:**

In conclusion, this analysis sheds light on global mental health trends, revealing complex dynamics. By employing rigorous data preprocessing, temporal analysis, and correlation assessments, nuanced patterns emerge. Predictive modeling using Random Forest Regressor shows promising accuracy in estimating depression prevalence. These insights highlight the urgency for tailored interventions to address mental health disparities globally, paving the way for targeted strategies and further research.

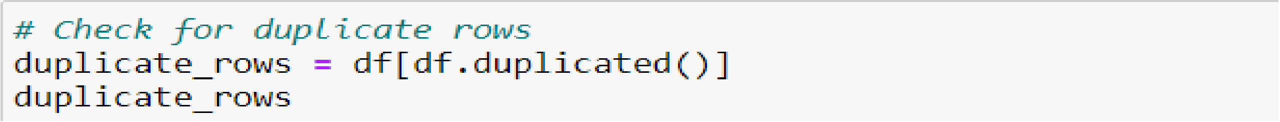
c. Diabetes Prediction

- **Code Link:** https://github.com/DharaKhamar/Healthcare-Pro--Diabetes-Prediction/tree/master
- **Dataset Used:**
  https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/code
  The Diabetes prediction dataset comprises medical and demographic data from patients, including diabetes status (positive/negative), age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. It's ideal for building ML models to predict diabetes and aid healthcare professionals in identifying at-risk patients and creating personalized treatment plans. Researchers can also utilize the dataset to investigate relationships between medical/demographic factors and diabetes likelihood.

- **Data Preprocessing:**
  Missing values check & Duplicate values check

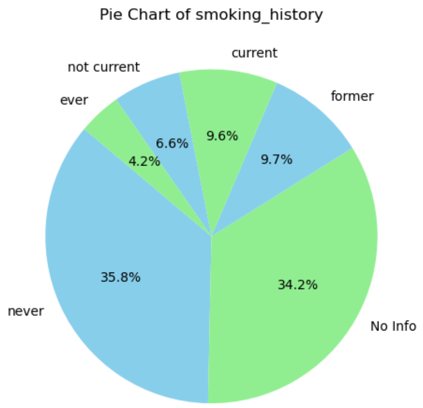


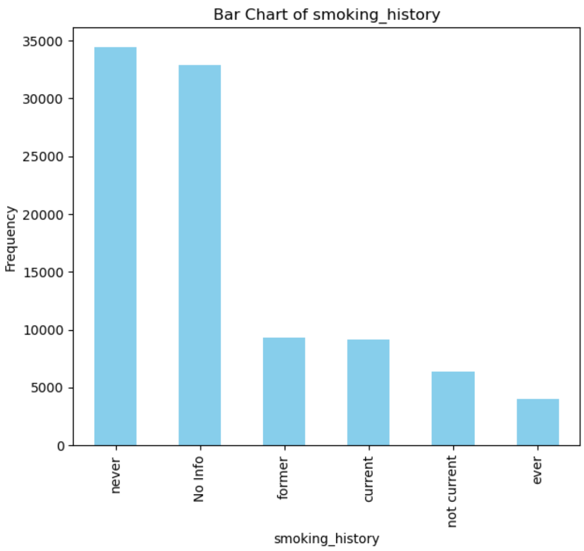
```
# Check for null values in each column
null_values_per_column = df.isnull().sum()
null_values_per_column
```

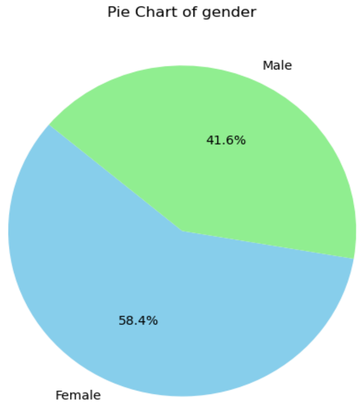


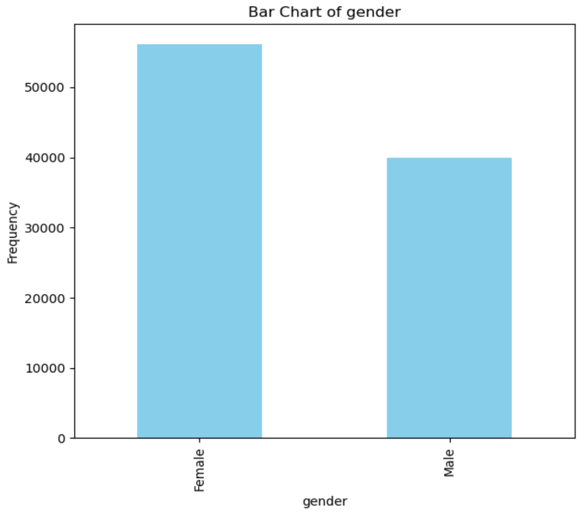
```
# Check for duplicate rows
duplicate_rows = df[df.duplicated()]
duplicate_rows
```

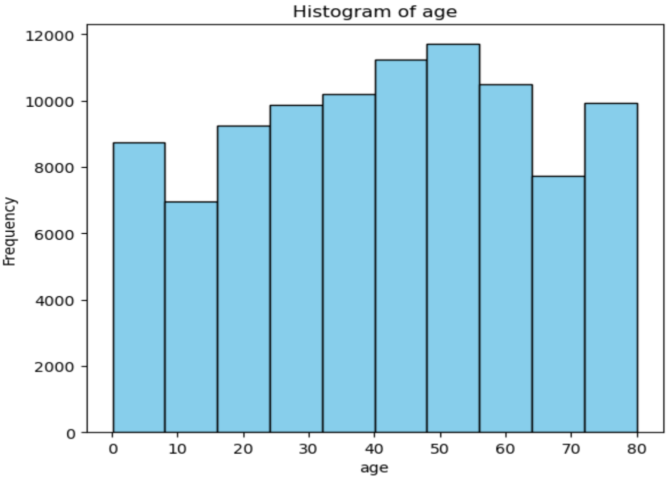
- **Exploratory Data Analysis (EDA):**

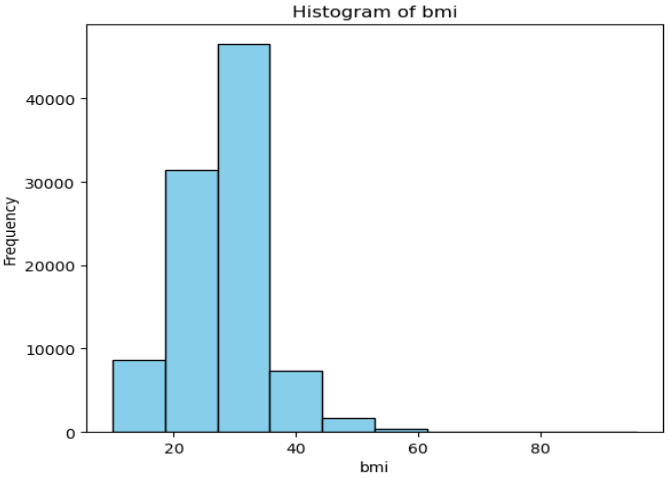


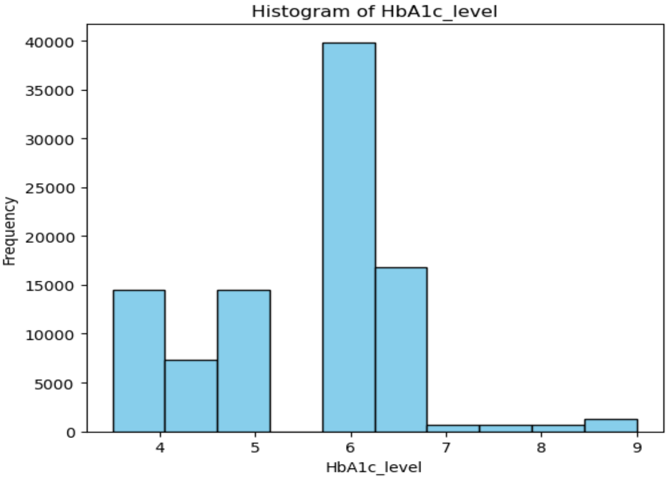


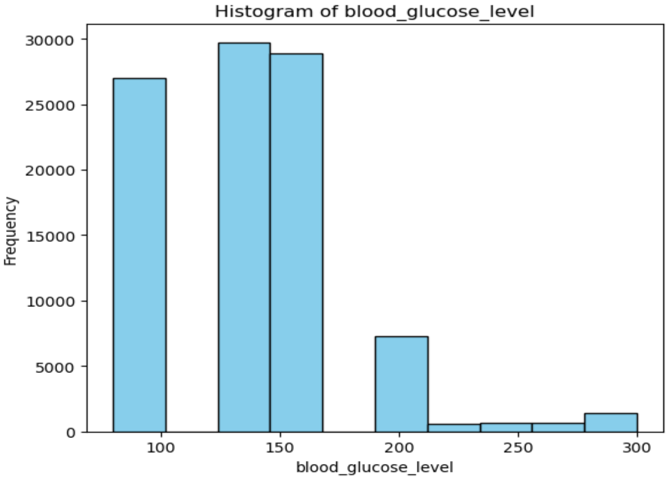


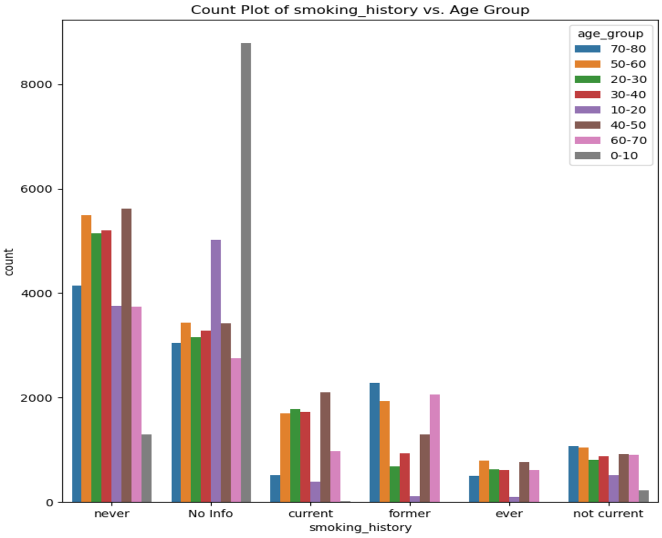


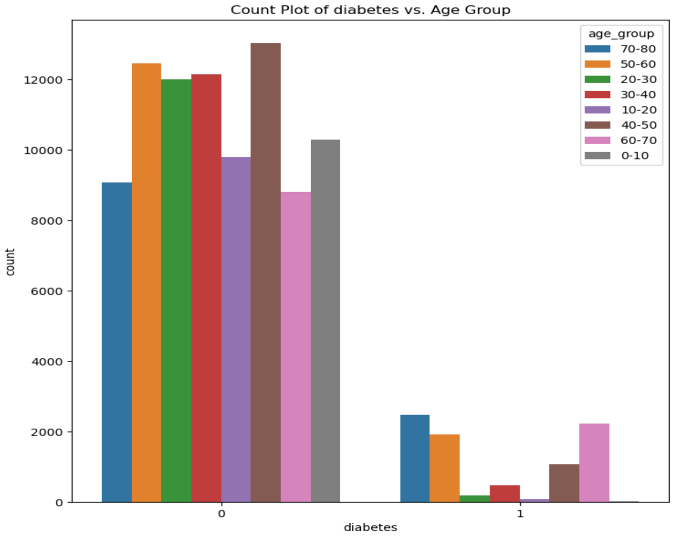

Histogram of age
Frequency
0
2000
4000
6000
8000
10000
12000
0
10
20
30
40
50
60
70
80
age

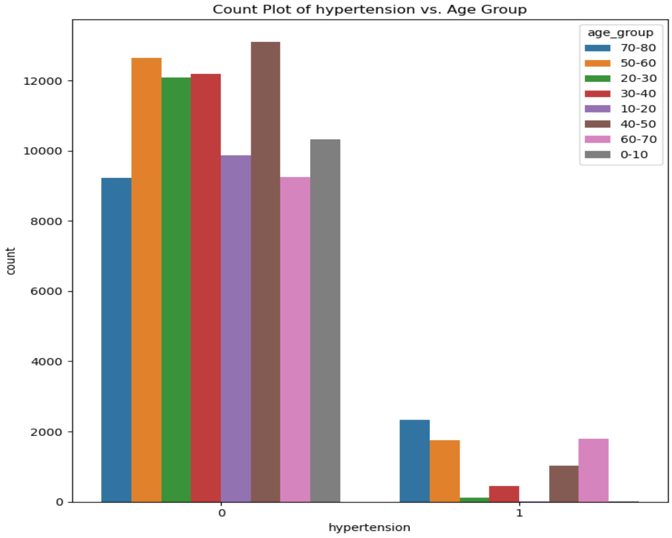

Histogram of bmi
Frequency
0
10000
20000
30000
40000
20
40
60
80
bmi

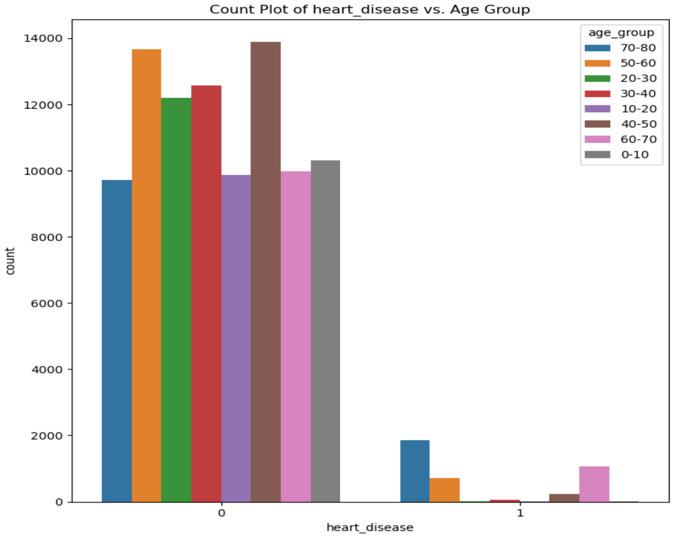

Histogram of HbA1c_level
Frequency
0
5000
10000
15000
20000
25000
30000
35000
40000
4
5
6
7
8
9
HbA1c_level

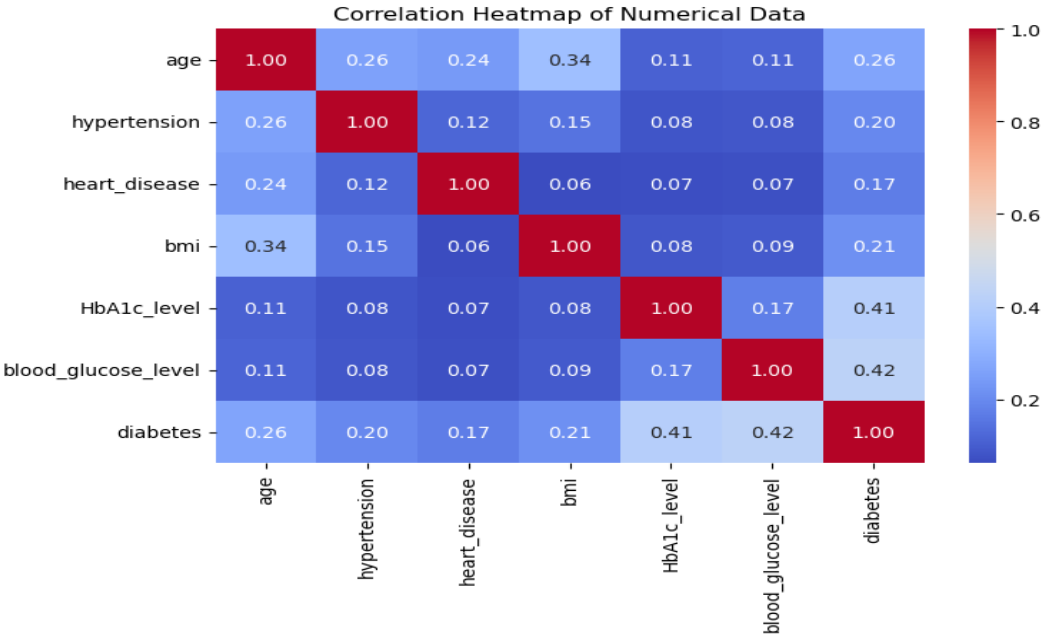

Histogram of blood_glucose_level
Frequency
0
5000
10000
15000
20000
25000
30000
100
150
200
250
300
blood_glucose_level


Count Plot of smoking_history vs. Age Group
count
0
2000
4000
6000
8000
never
No Info
current
former
ever
not current
smoking_history
age_group
70-80
50-60
20-30
30-40
10-20
40-50
60-70
0-10

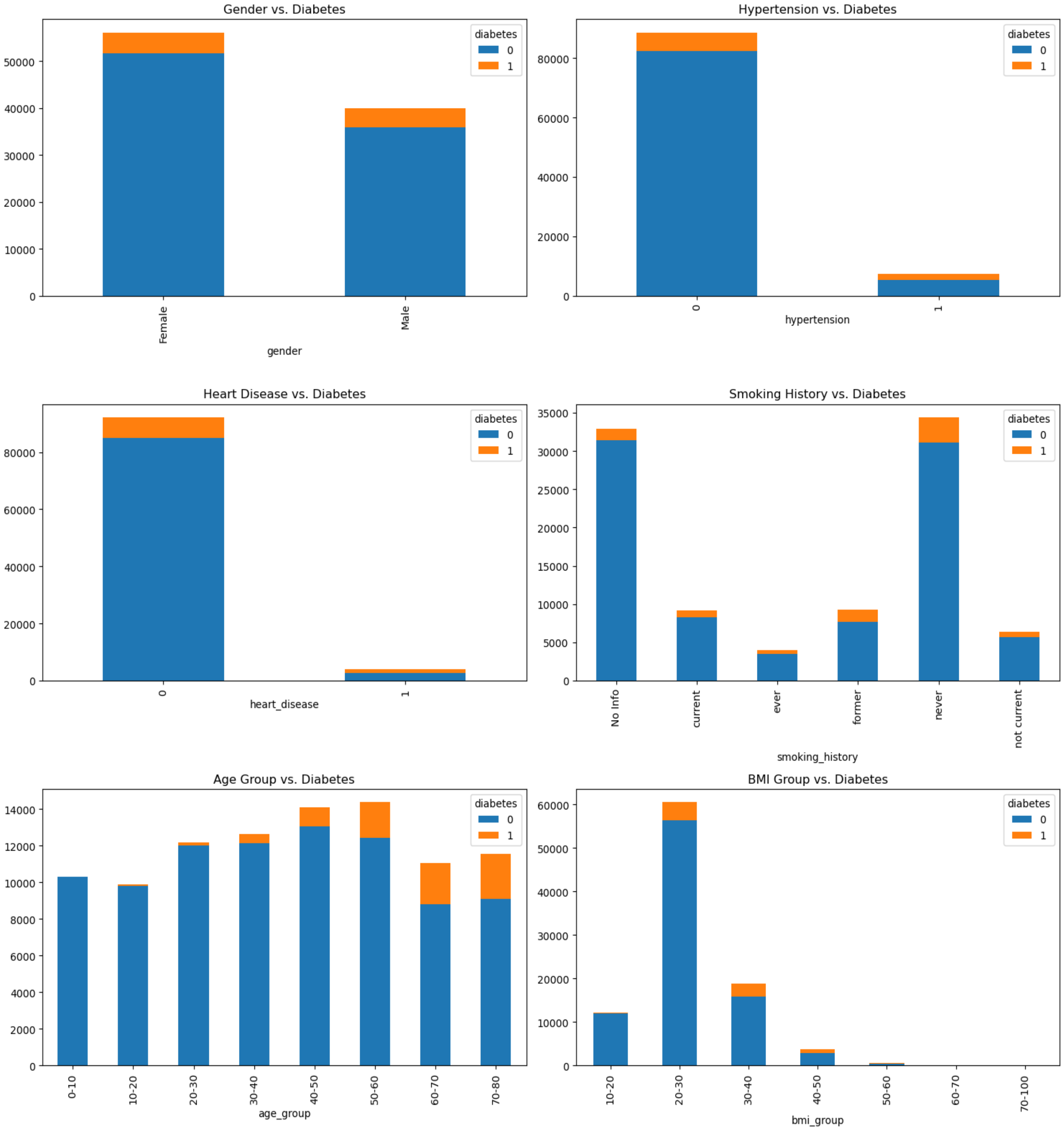

Count Plot of diabetes vs. Age Group
count
0
2000
4000
6000
8000
10000
12000
0
1
diabetes
age_group
70-80
50-60
20-30
30-40
10-20
40-50
60-70
0-10


Count Plot of hypertension vs. Age Group
count
0
2000
4000
6000
8000
10000
12000
0
1
hypertension
age_group
70-80
50-60
20-30
30-40
10-20
40-50
60-70
0-10


Count Plot of heart_disease vs. Age Group
count
0
2000
4000
6000
8000
10000
12000
14000
0
1
heart_disease
age_group
70-80
50-60
20-30
30-40
10-20
40-50
60-70
0-10



Correlation Heatmap of Numerical Data



Gender vs. Diabetes

Hypertension vs. Diabetes

Heart Disease vs. Diabetes

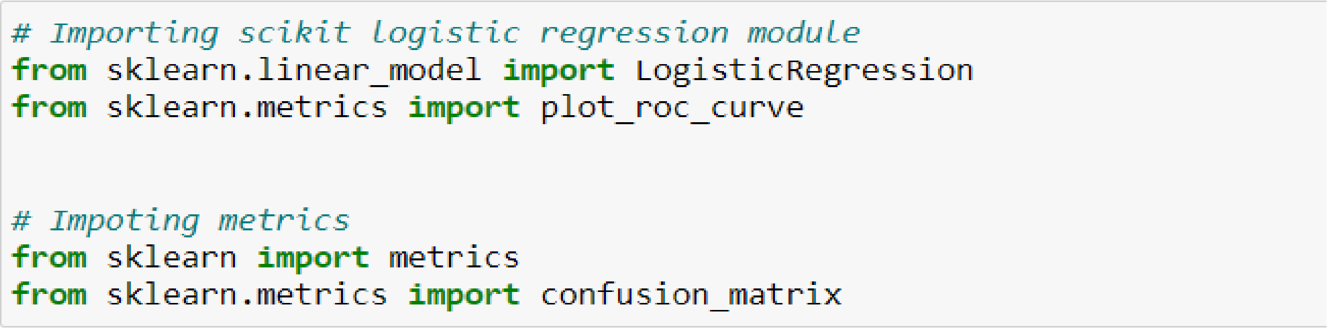
Smoking History vs. Diabetes

Age Group vs. Diabetes

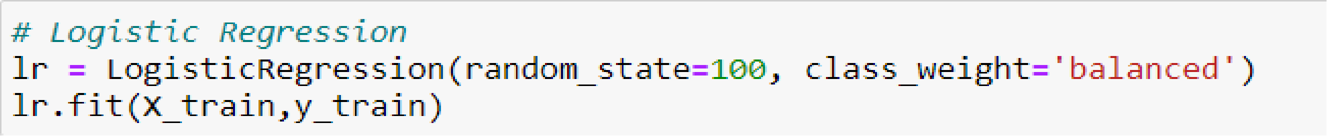
BMI Group vs. Diabetes

➢ Model Building:

○ Logistic regression

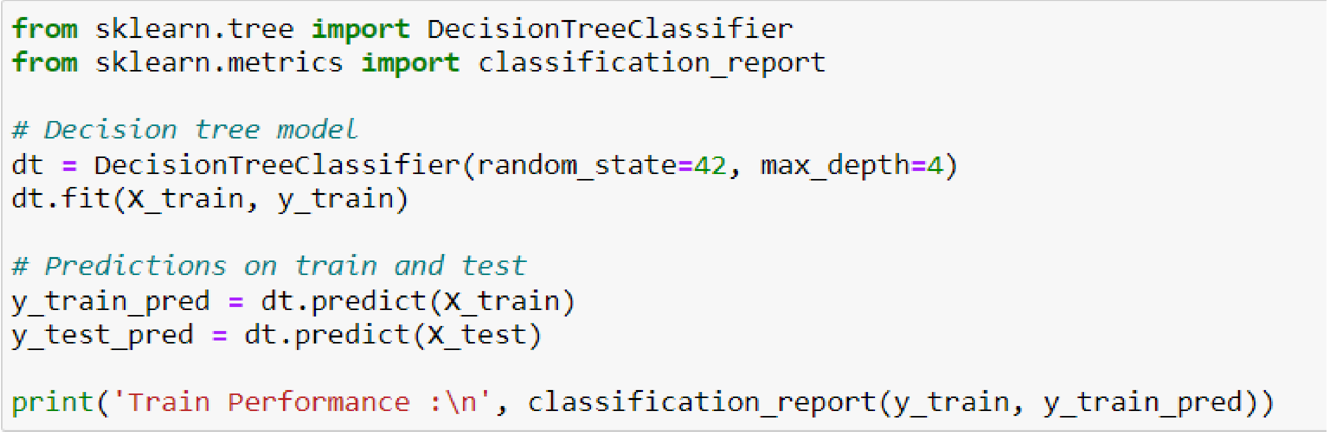


```
# Importing scikit logistic regression module
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import plot_roc_curve


# Impoting metrics
from sklearn import metrics
from sklearn.metrics import confusion_matrix
```



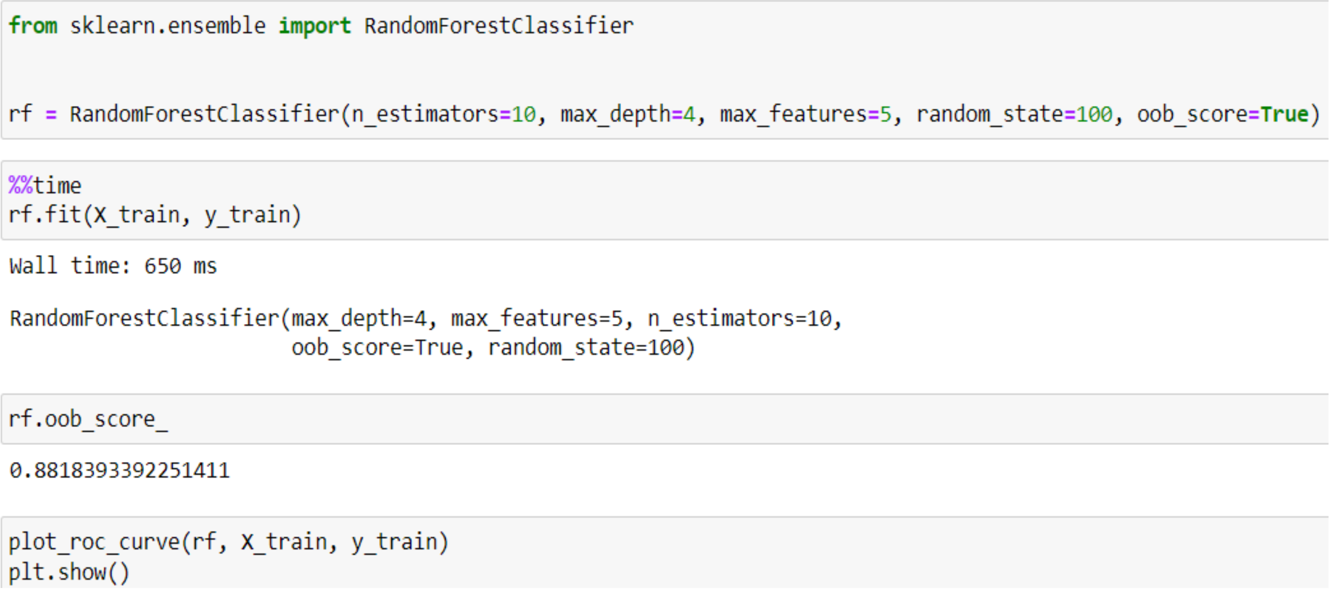
```
# Logistic Regression
lr = LogisticRegression(random_state=100, class_weight='balanced')
lr.fit(X_train,y_train)
```

○ Decision Tree



```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report

# Decision tree model
dt = DecisionTreeClassifier(random_state=42, max_depth=4)
dt.fit(X_train, y_train)

# Predictions on train and test
y_train_pred = dt.predict(X_train)
y_test_pred = dt.predict(X_test)

print('Train Performance :\n', classification_report(y_train, y_train_pred))
```

○ Random Forest



```
from sklearn.ensemble import RandomForestClassifier


rf = RandomForestClassifier(n_estimators=10, max_depth=4, max_features=5, random_state=100, oob_score=True)
```

```
%%time
rf.fit(X_train, y_train)
```

```
Wall time: 650 ms

RandomForestClassifier(max_depth=4, max_features=5, n_estimators=10,
                       oob_score=True, random_state=100)
```

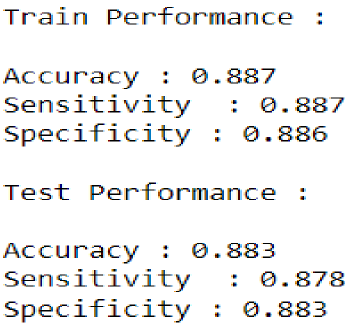
```
rf.oob_score_
```

```
0.8818393392251411
```

```
plot_roc_curve(rf, X_train, y_train)
plt.show()
```

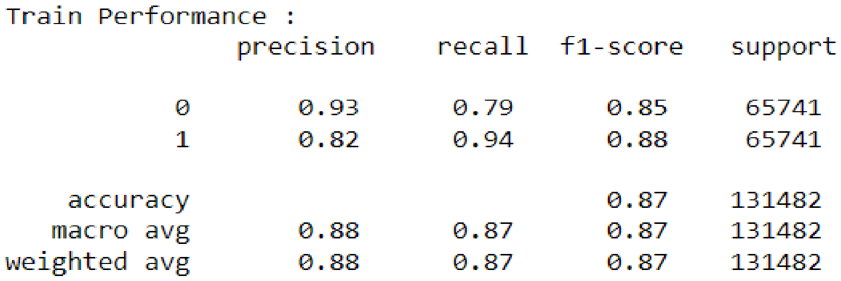
- **Evaluation:**

  - **Logistic Regression:**

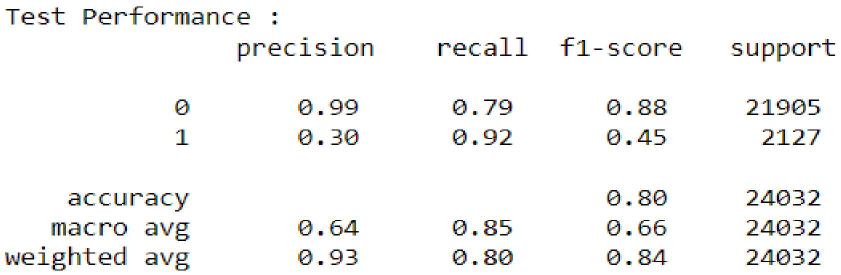


```
Train Performance :

Accuracy : 0.887
Sensitivity  : 0.887
Specificity : 0.886

Test Performance :

Accuracy : 0.883
Sensitivity  : 0.878
Specificity : 0.883
```

  - **Decision Tree:**

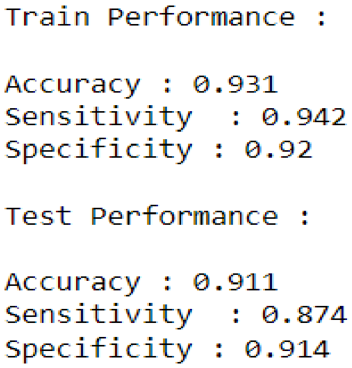


```
Train Performance :
              precision    recall  f1-score   support

           0       0.93      0.79      0.85     65741
           1       0.82      0.94      0.88     65741

    accuracy                           0.87    131482
   macro avg       0.88      0.87      0.87    131482
weighted avg       0.88      0.87      0.87    131482
```



```
Test Performance :
              precision    recall  f1-score   support

           0       0.99      0.79      0.88     21905
           1       0.30      0.92      0.45      2127

    accuracy                           0.80     24032
   macro avg       0.64      0.85      0.66     24032
weighted avg       0.93      0.80      0.84     24032
```

  - **Random Forest:**



```
Train Performance :

Accuracy : 0.931
Sensitivity  : 0.942
Specificity : 0.92

Test Performance :

Accuracy : 0.911
Sensitivity  : 0.874
Specificity : 0.914
```

- **Conclusion:**

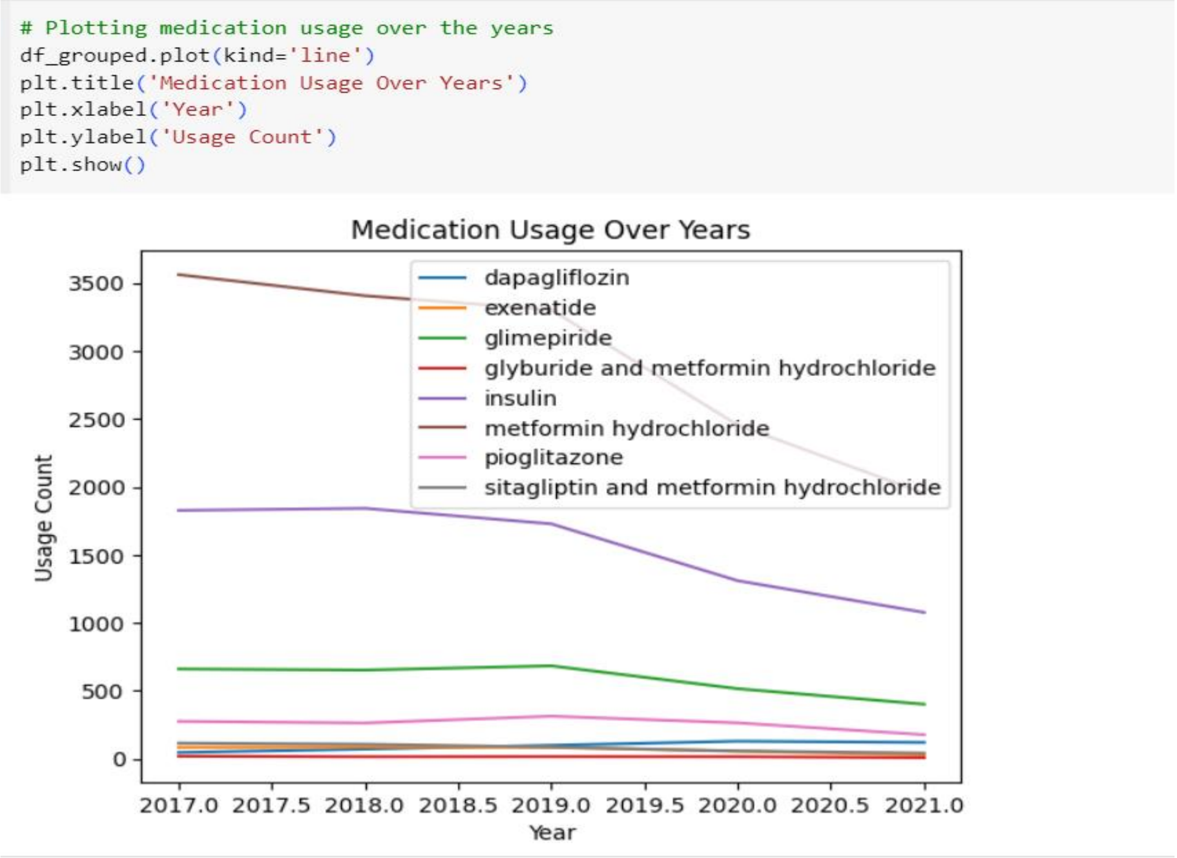
In conclusion, the dataset reveals a predominance of females (58%) and a significant proportion of non-smokers (35%), with missing smoking history data for 34% of individuals. Most lack hypertension (92%) or heart disease (96%). Age-wise, the data is concentrated between 40-60 years, with diabetes prevalence notably increasing after 40 and particularly after 60. Additionally, BMI, HbA1c_level, and blood glucose levels predominantly fall within specific ranges. However, there is a notable data imbalance, with only 8.8% having diabetes. Random Forest demonstrates the highest accuracy (91%) among tested models, making it the preferred choice for diabetes prediction.

8. Financial Modelling (equation) with Machine Learning & Data Analysis

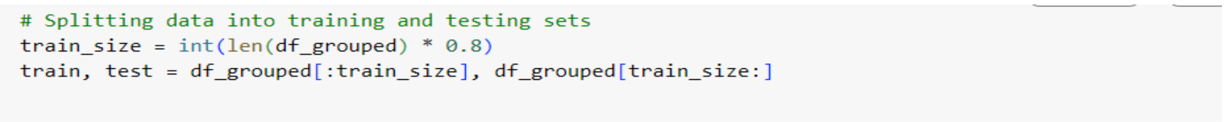
Dataset link : https://www.kaggle.com/datasets/i191796majid/cost-of-diabetes-in-usa
The dataset contains information on medication usage related to diabetes management over multiple years. Each row represents a specific year, and the columns indicate the usage of various medications commonly prescribed for diabetes treatment, such as dapagliflozin, exenatide, glimepiride, glyburide and metformin hydrochloride, insulin, metformin hydrochloride, pioglitazone, and sitagliptin and metformin hydrochloride. Medication usage is represented as binary values (0 indicating no usage, 1 indicating usage) for each medication in each year.

Dataset link : https://www.kaggle.com/datasets/rifkaregmi/usa-mental-health-dataset

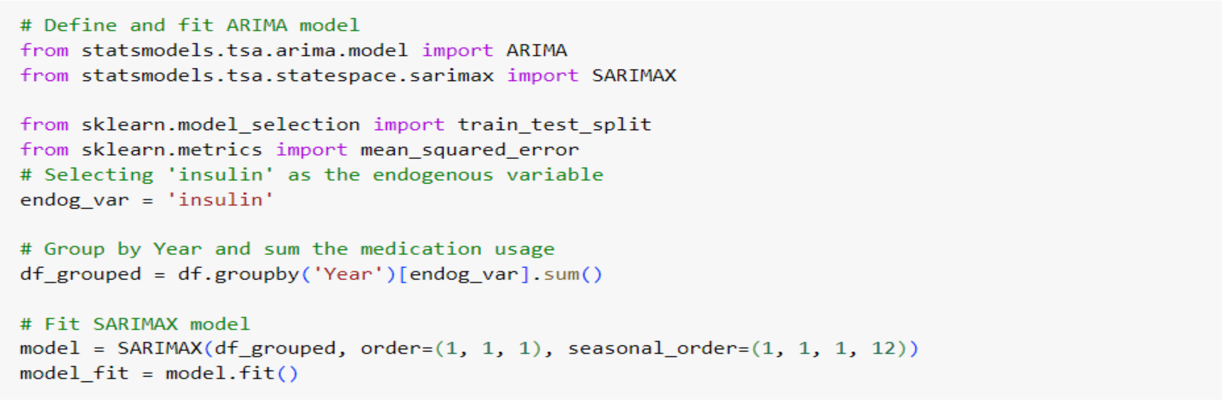
The dataset contains information related to mental health prevalence in Nevada, USA, spanning the years 2018 and 2019. It includes data on the prevalence of recent mentally unhealthy days among adults, with stratification by gender and race/ethnicity. Additionally, the dataset provides confidence intervals for the reported prevalence rates.



```
# Plotting medication usage over the years
df_grouped.plot(kind='line')
plt.title('Medication Usage Over Years')
plt.xlabel('Year')
plt.ylabel('Usage Count')
plt.show()
```

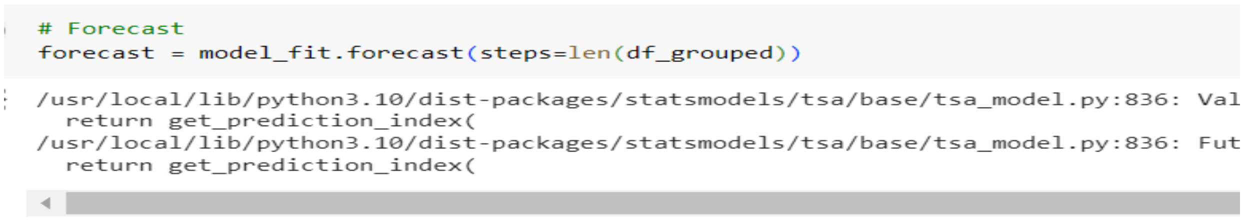


```
# Splitting data into training and testing sets
train_size = int(len(df_grouped) * 0.8)
train, test = df_grouped[:train_size], df_grouped[train_size:]
```



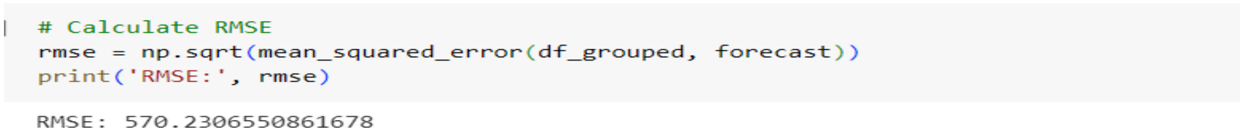
```
# Define and fit ARIMA model
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.statespace.sarimax import SARIMAX

from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# Selecting 'insulin' as the endogenous variable
endog_var = 'insulin'

# Group by Year and sum the medication usage
df_grouped = df.groupby('Year')[endog_var].sum()

# Fit SARIMAX model
model = SARIMAX(df_grouped, order=(1, 1, 1), seasonal_order=(1, 1, 1, 12))
model_fit = model.fit()
```

- Plots the usage of each medication over time.
- Performs seasonal decomposition on the 'metformin hydrochloride' usage to understand its trend, seasonal, and residual components.
- Forecast the 'metformin hydrochloride' usage for the next few years using the ARIMA model.



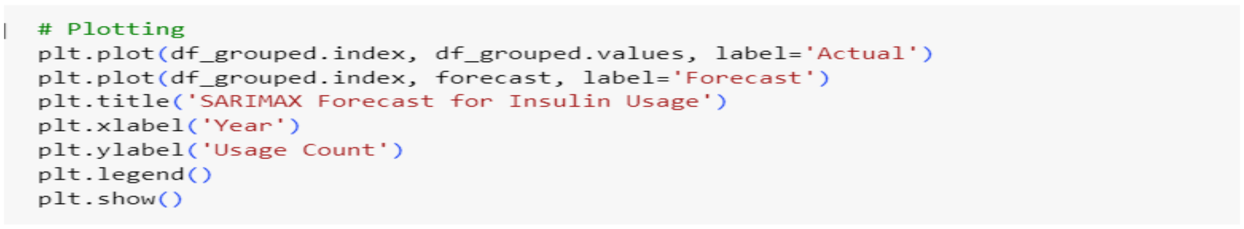
```
# Forecast
forecast = model_fit.forecast(steps=len(df_grouped))
```

```
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/base/tsa_model.py:836: Val
  return get_prediction_index(
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/base/tsa_model.py:836: Fut
  return get_prediction_index(
```

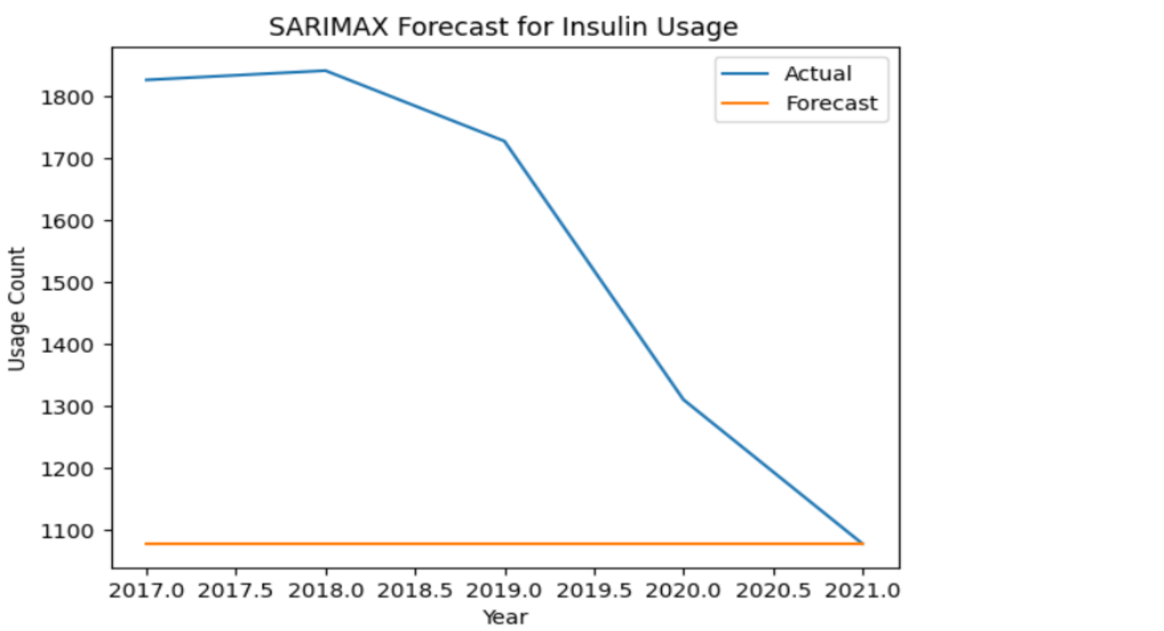


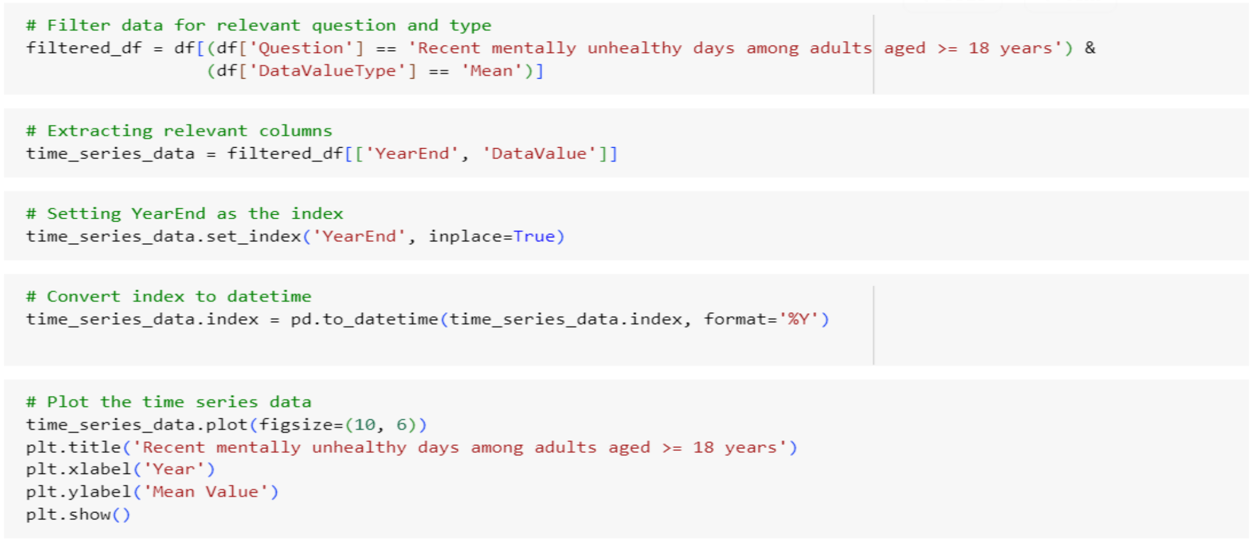
```
# Calculate RMSE
rmse = np.sqrt(mean_squared_error(df_grouped, forecast))
print('RMSE:', rmse)
```

```
RMSE:  570.2306550861678
```



```
# Plotting
plt.plot(df_grouped.index, df_grouped.values, label='Actual')
plt.plot(df_grouped.index, forecast, label='Forecast')
plt.title('SARIMAX Forecast for Insulin Usage')
plt.xlabel('Year')
plt.ylabel('Usage Count')
plt.legend()
plt.show()
```





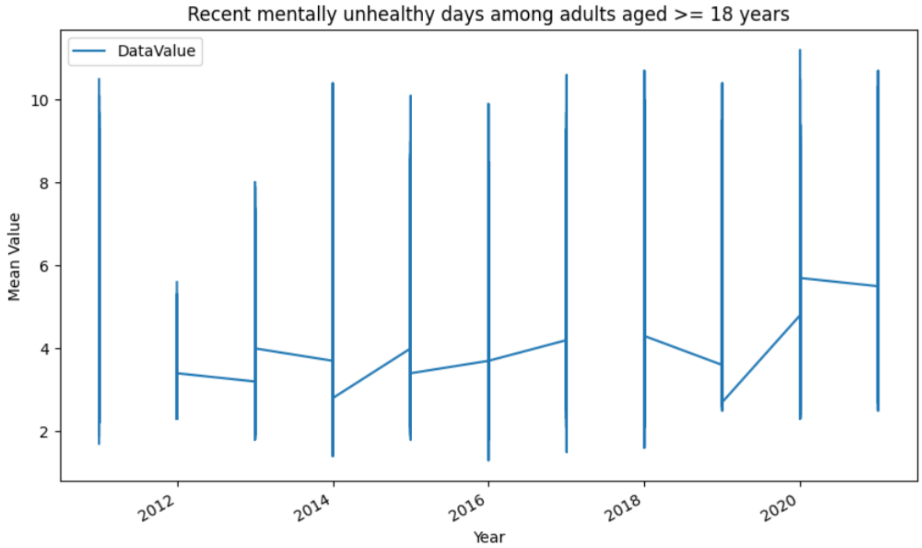
```
# Filter data for relevant question and type
filtered_df = df[(df['Question'] == 'Recent mentally unhealthy days among adults aged >= 18 years') &
                 (df['DataValueType'] == 'Mean')]
```

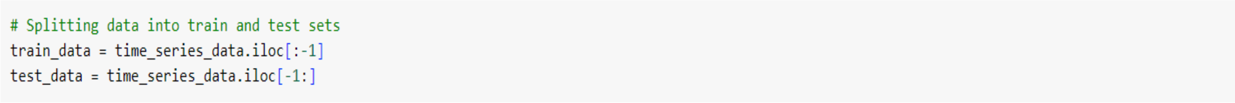
```
# Extracting relevant columns
time_series_data = filtered_df[['YearEnd', 'DataValue']]
```

```
# Setting YearEnd as the index
time_series_data.set_index('YearEnd', inplace=True)
```

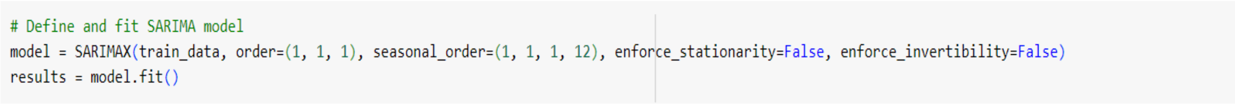
```
# Convert index to datetime
time_series_data.index = pd.to_datetime(time_series_data.index, format='%Y')
```

```
# Plot the time series data
time_series_data.plot(figsize=(10, 6))
plt.title('Recent mentally unhealthy days among adults aged >= 18 years')
plt.xlabel('Year')
plt.ylabel('Mean Value')
plt.show()
```

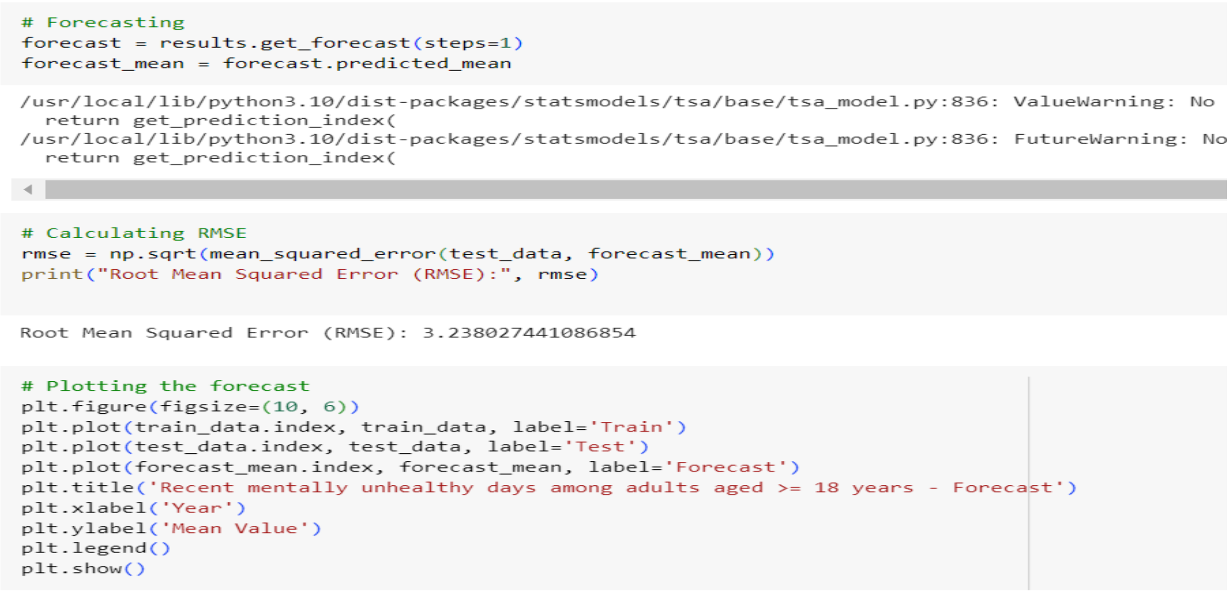




```
# Splitting data into train and test sets
train_data = time_series_data.iloc[:-1]
test_data = time_series_data.iloc[-1:]
```



```
# Define and fit SARIMA model
model = SARIMAX(train_data, order=(1, 1, 1), seasonal_order=(1, 1, 1, 12), enforce_stationarity=False, enforce_invertibility=False)
results = model.fit()
```



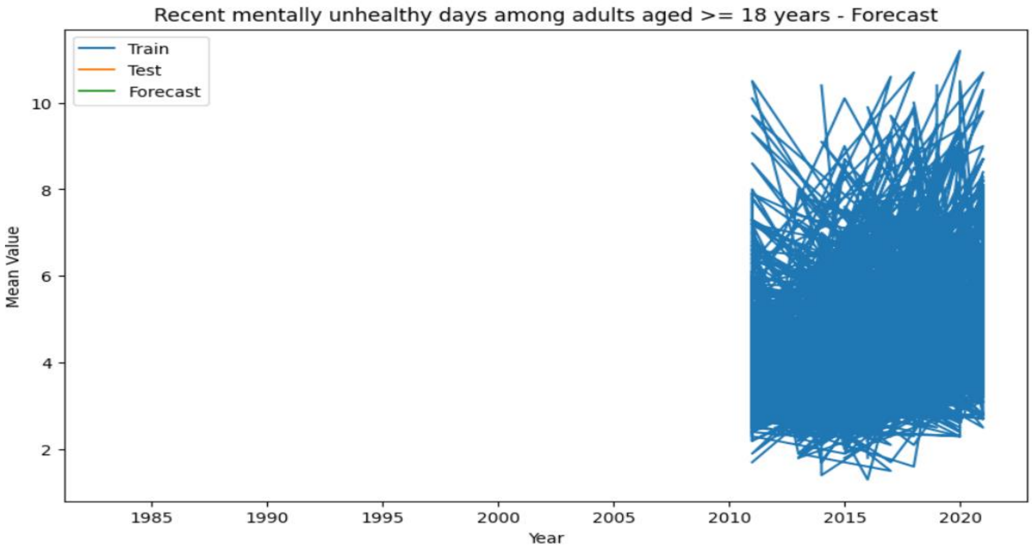
```
# Forecasting
forecast = results.get_forecast(steps=1)
forecast_mean = forecast.predicted_mean
```

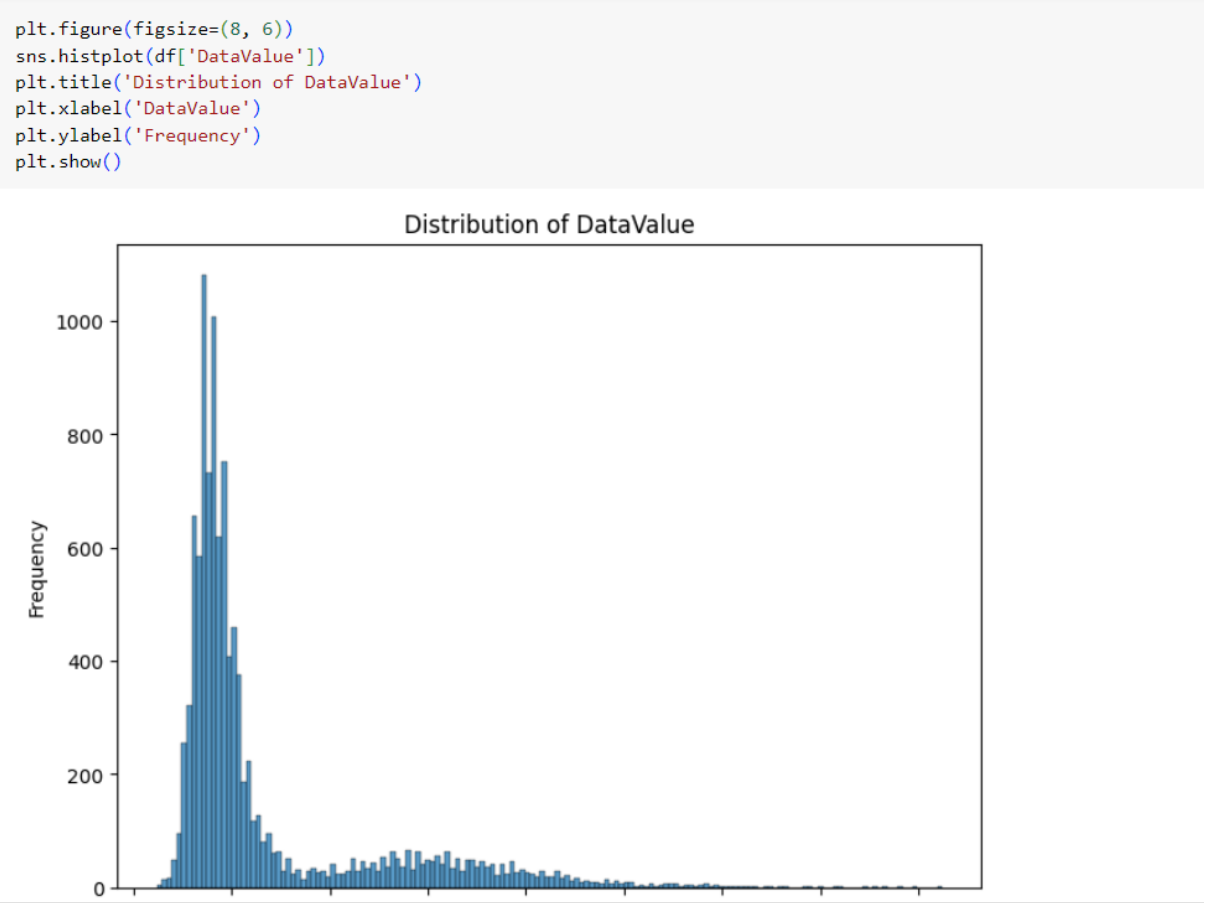
```
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/base/tsa_model.py:836: ValueWarning: No
  return get_prediction_index(
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/base/tsa_model.py:836: FutureWarning: No
  return get_prediction_index(
```

```
# Calculating RMSE
rmse = np.sqrt(mean_squared_error(test_data, forecast_mean))
print("Root Mean Squared Error (RMSE):", rmse)
```

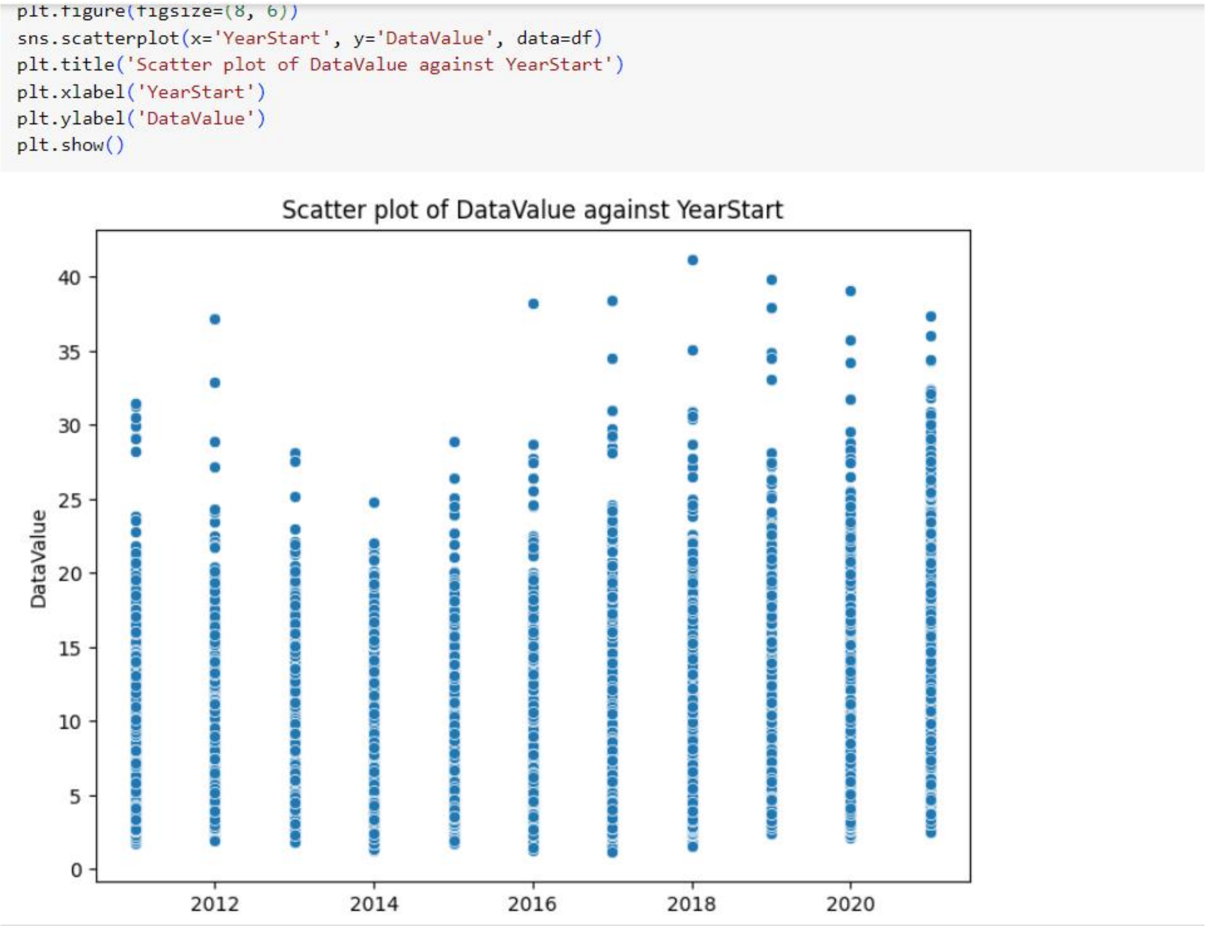
```
Root Mean Squared Error (RMSE): 3.238027441086854
```

```
# Plotting the forecast
plt.figure(figsize=(10, 6))
plt.plot(train_data.index, train_data, label='Train')
plt.plot(test_data.index, test_data, label='Test')
plt.plot(forecast_mean.index, forecast_mean, label='Forecast')
plt.title('Recent mentally unhealthy days among adults aged >= 18 years - Forecast')
plt.xlabel('Year')
plt.ylabel('Mean Value')
plt.legend()
plt.show()
```

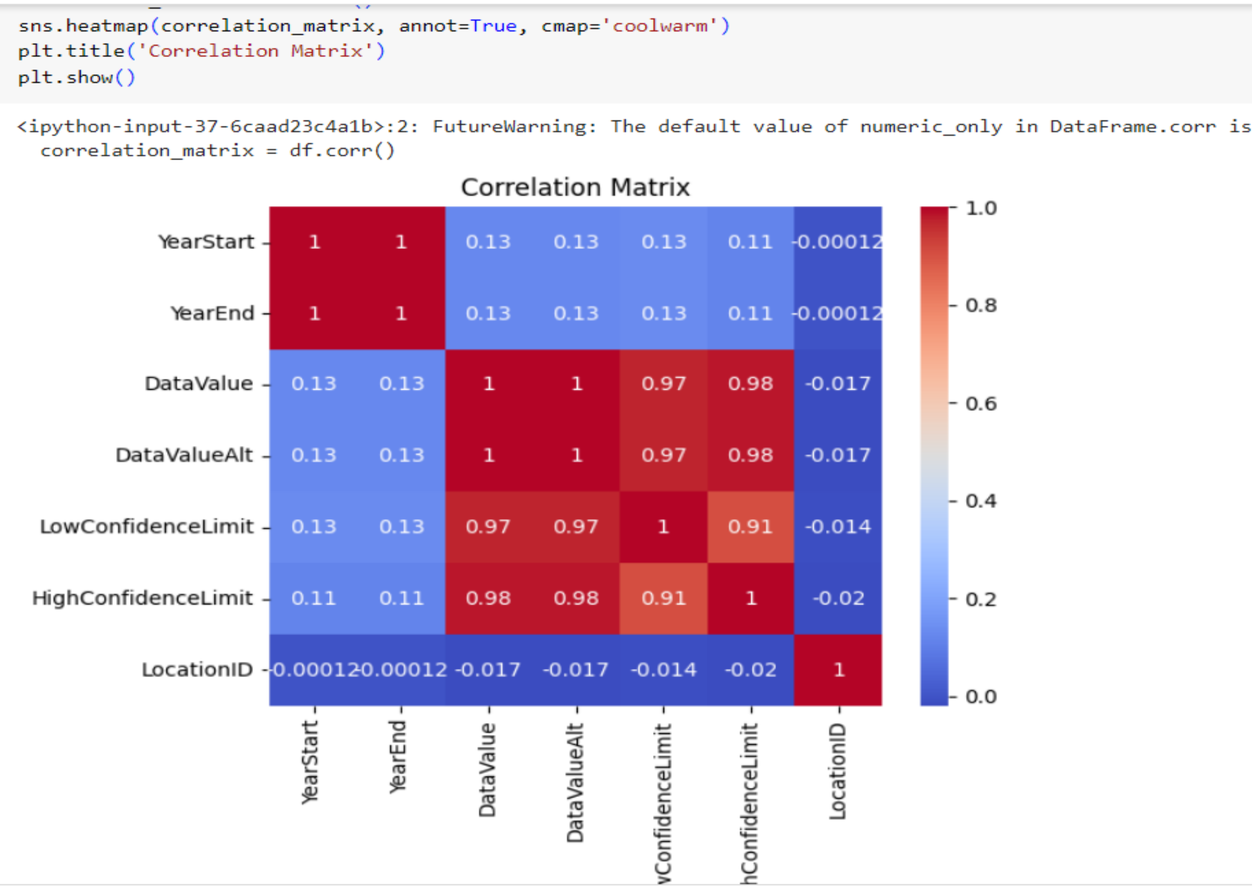




```
plt.figure(figsize=(8, 6))
sns.histplot(df['DataValue'])
plt.title('Distribution of DataValue')
plt.xlabel('DataValue')
plt.ylabel('Frequency')
plt.show()
```



```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='YearStart', y='DataValue', data=df)
plt.title('Scatter plot of DataValue against YearStart')
plt.xlabel('YearStart')
plt.ylabel('DataValue')
plt.show()
```



```
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
<ipython-input-37-6caad23c4a1b>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is
  correlation_matrix = df.corr()
```

In 2024, India's healthcare IT sector is witnessing significant growth, with AI technology playing a central role.

Momentum Areas & Emerging Trends:

1. AI in diagnostics and decision support.
2. Telemedicine and remote monitoring.
3. Data privacy and security.
4. Personalized medicine and AI-driven trials.
5. Wearables and assistive tech.
6. Mental health support with AI.

Integration of AI: AI is enhancing various aspects of healthcare, from diagnostics to telemedicine, offering streamlined solutions.

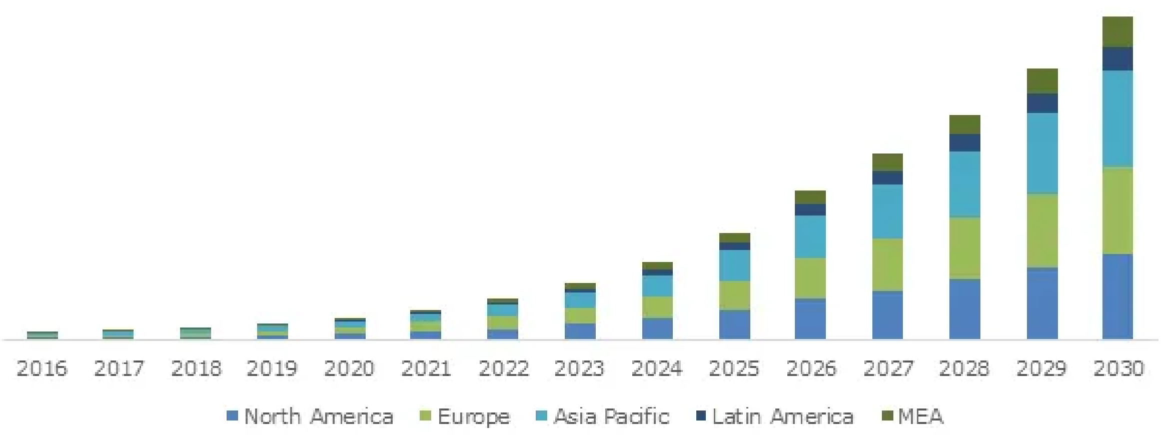
Government Initiatives: Programs like the Ayushman Bharat Digital Mission are driving digital health adoption, creating health IDs and digitizing records.

Market Growth and Cost Savings: Healthcare tech adoption is expected to grow the market significantly by 2030, driven by AI's cost-saving potential.

Voice AI's Potential: Voice biomarkers hold promise for early disease detection, particularly beneficial in regions with limited healthcare access.

In 2023, the rise of mobile health (mHealth) stands out as a key trend in healthcare app development. Leveraging mobile technology, these apps offer personalized care, remote monitoring, and educational resources, enhancing patient-provider communication and engagement. They enable quick access to medical data, improving decision-making for healthcare professionals. Additionally, mHealth apps promote transparency, allowing patients to review records and communicate with their providers easily. With their transformative potential, mHealth apps are driving innovation in healthcare delivery.

Global mHealth Market, By Region, 2016 - 2030 (USD Million)



Source: www.gminsights.com

The Covid-19 pandemic accelerated the growth of India's health app market, driven by lifestyle diseases and increased mobile technology adoption. In 2021, health and fitness apps surged, with mental health apps like Calm seeing significant downloads. Popular apps like 1mg and Practo paved the way for traditional and emerging health-focused businesses. The market is expected to reach ₹337.89 billion by 2026. To succeed, developers need analytics platforms like Adjust for marketing metrics and data privacy compliance. Personalization is crucial for engaging users in their health journey.

Subscription Revenue: This is the primary source of income for subscription-based mobile applications. You'll need to consider the pricing strategy and the number of subscribers at each pricing tier.

$$\text{Subscription Revenue} = \sum_{i=1}^{n}(\text{Number of Subscribers}_i \times \text{Subscription Price}_i)$$

Where:

n is the number of subscription tiers.

$\text{Number of Subscribers}_i$ is the number of subscribers at subscription tier i.

$\text{Subscription Price}_i$ is the price of subscription tier i.

Growth Rate: The growth rate represents the percentage increase in revenue over a specific period. It accounts for factors such as user acquisition, market expansion, and product improvements.

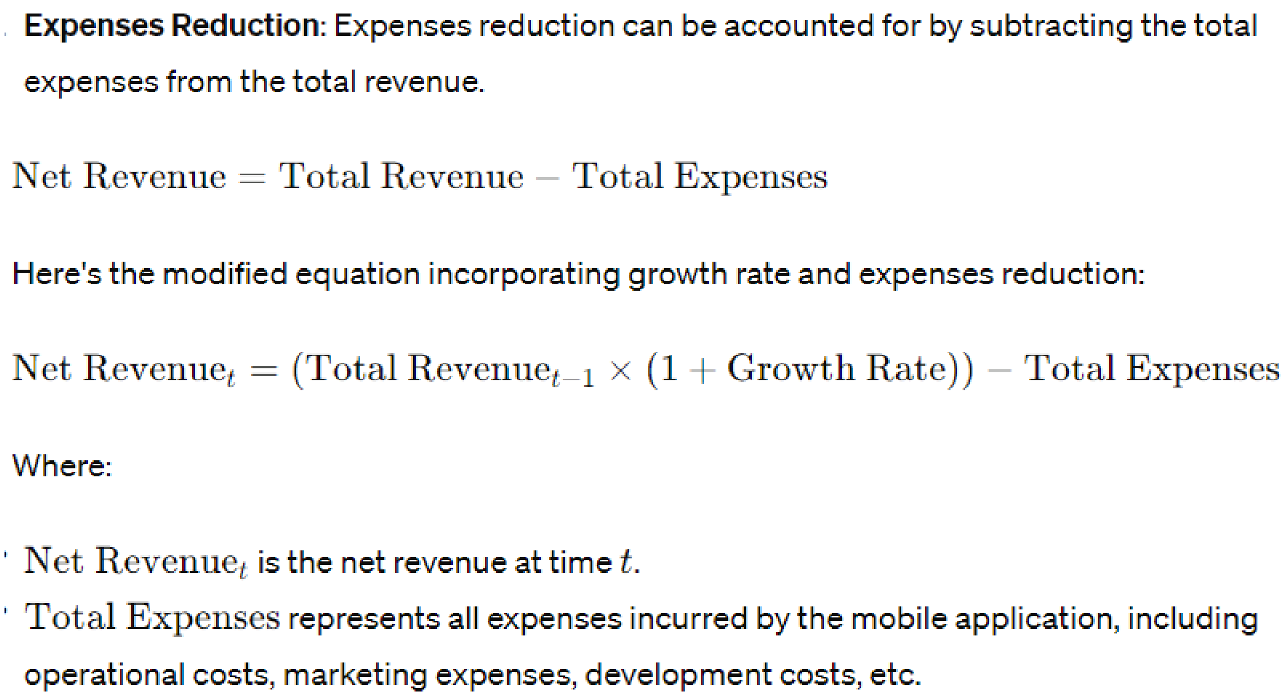
$$\text{Total Revenue}_t = \text{Total Revenue}_{t-1} \times (1 + \text{Growth Rate})$$

Where:

Total Revenue_t is the total revenue at time t.

$\text{Total Revenue}_{t-1}$ is the total revenue at the previous time period.

Growth Rate is the percentage increase in revenue.



- **Expenses Reduction:** Expenses reduction can be accounted for by subtracting the total expenses from the total revenue.

$$\text{Net Revenue} = \text{Total Revenue} - \text{Total Expenses}$$

Here's the modified equation incorporating growth rate and expenses reduction:

$$\text{Net Revenue}_t = (\text{Total Revenue}_{t-1} \times (1 + \text{Growth Rate})) - \text{Total Expenses}$$

Where:

- Net Revenue_t is the net revenue at time t.
- Total Expenses represents all expenses incurred by the mobile application, including operational costs, marketing expenses, development costs, etc.

9. Conclusion:

- Our analysis of global mental health trends and chronic conditions highlights crucial insights for our healthcare mobile application. Through rigorous data analysis and predictive modeling, we've identified nuanced patterns and correlations, emphasizing the need for tailored interventions.
- With a focus on addressing mental health disparities and chronic diseases, our application aims to provide personalized support to diverse user demographics. Despite data challenges, our chosen machine learning models, particularly Random Forest, have shown promising accuracy rates.
- Moving forward, our application will leverage these insights to deliver targeted interventions and foster healthier outcomes worldwide.