Personalized Healthcare Recommendations

About the Dataset (given information):

Blood datasets typically encompass a broad array of information related to hematology, blood chemistry, and related health indicators. These datasets include datapoints such as blood cell counts, hemoglobin levels, hematocrit, platelet counts, white blood cell differentials, and various blood chemistry parameters such as glucose, cholesterol, and electrolyte levels.

These datasets are invaluable for medical research, clinical diagnostics, and public health initiatives. Researchers and healthcare professionals utilize blood datasets to study hematological disorders, monitor disease progression, assess treatment efficacy, and identify risk factors for various health conditions.

Machine learning techniques are often applied to blood datasets to develop predictive models for diagnosing diseases, predicting patient outcomes, and identifying biomarkers associated with specific health conditions. These models can assist clinicians in making more accurate diagnoses, designing personalized treatment plans, and improving patient care.

Additionally, blood datasets play a crucial role in epidemiological studies and population health research. By analyzing large-scale blood datasets, researchers can identify trends in blood parameters across different demographic groups, and assess the prevalence of blood disorders, and evaluate the impact of lifestyle factors and environmental exposures on hematological health.

Overall, blood datasets serve as valuable resources for advancing our understanding of hematology, improving healthcare practices, and promoting better health.

Personalized Healthcare Recommendations Machine Learning Project:

Project Overview (given information**):**

The Personalized Healthcare Recommendations project aims to develop a machine learning model that provides tailored healthcare recommendations based on individual patient data. This can include recommendations for lifestyle changes, preventive measures, medications, or treatment plans. The goal is to improve patient outcomes by leveraging data-driven insights to offer personalized advice.

Project Steps:

1. Understanding the Problem:

○ The goal is to provide personalized healthcare recommendations to patients based on their health data, medical history, lifestyle, and other relevant factors.
○ Use machine learning techniques to analyze patient data and generate actionable insights

2. Dataset Preparation:

○ Data Sources: Collect data from various sources such as electronic health records (EHRs), wearable devices, patient surveys, and publicly available health datasets.
○ Features: Include demographic information (age, gender), medical history, lifestyle factors (diet, exercise), biometric data (blood pressure, heart rate), lab results, and medication history.
○ Labels: Recommendations or health outcomes (if available).

3. Data Exploration and Visualization:
○ Load and explore the dataset using descriptive statistics and visualization techniques.
○ Use libraries like Pandas for data manipulation and Matplotlib/Seaborn for visualization.
○ Identify patterns, correlations, and distributions in the data.

4. Data Preprocessing:
○ Handle missing values through imputation or removal.
○ Standardize or normalize continuous features.
○ Encode categorical variables using techniques like one-hot encoding.
○ Split the dataset into training, validation, and testing sets.

5. Feature Engineering:
○ Create new features that may be useful for prediction, such as health indices or composite scores.
○ Perform feature selection to identify the most relevant features for the model.

6. Model Selection and Training:
○ Choose appropriate machine learning algorithms based on the problem.

Common choices include:

■ Logistic Regression
■ Decision Trees
■ RandomForest
■ Gradient Boosting Machines (e.g., XGBoost)
■ Support Vector Machine (SVM)
■ Neural Networks
○ Train multiple models to find the best-performing one.

7. Model Evaluation:
○ Evaluate the models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.
○ Use cross-validation to ensure the model generalizes well to unseen data.
○ Visualize model performance using confusion matrices, ROC curves, and other relevant plots.

8. Recommendation System Implementation:
○ Develop an algorithm to generate personalized recommendations based on the model's predictions.
○ Use techniques like collaborative filtering or content-based filtering if incorporating user feedback or preferences.
○ Ensure recommendations are interpretable and actionable for healthcare professionals and patients

Acknowledgements: The data was collected from the National Health and Nutrition Examination Survey (NHANES) through the Kaggle website.

Project Overview (Final Report)

This project aims to build a **machine learning system** that provides **tailored healthcare recommendations** using patient data such as age, vitals, lab results, and lifestyle factors. The solution is designed to assist healthcare professionals by highlighting key predictors of health risk and suggesting actionable advice for patients.

1. Problem Understanding

We aim to:

- Predict and classify patient health risks.
- Offer personalized, interpretable recommendations based on medical indicators.
- Use patient-level data from blood work, examination records, demographics, and lifestyle surveys.

2. Dataset Preparation

- Data was compiled from multiple sources:
    - `blood.csv`, `labs.csv`, `examination.csv`, `medications.csv`, `questionnaire.csv`, `demographic.csv`, `diet.csv`
- All files were merged on the common key `SEQN`.
- Each dataset contains medical features relevant to patient diagnosis and wellness tracking.

3. Data Exploration & Visualization

- Explored numerical distributions (e.g., Age, Cholesterol).
- Identified and visualized correlations among features using a heatmap.
- Plotted box plots to observe relationships between health indicators and recommendation labels.

4. Data Preprocessing

- Handled missing values by dropping incomplete records relevant to selected features.
- Encoded categorical features using OneHotEncoding.
- Scaled numerical features using `StandardScaler`.

5. Feature Engineering

- Selected predictive variables like blood pressure, BMI, cholesterol, and age.
- Combined lab results with patient history and lifestyle indicators.
- Created a new feature: `recommendation` score (levels 0–3).

6. Model Training

- Trained a `RandomForestClassifier` inside a `Pipeline`.
- Used both numeric and categorical pipelines within `ColumnTransformer`.

7. Model Evaluation

- Evaluated predictions using:
    - Confusion Matrix
    - Precision, Recall, F1-score
- Visualized top features contributing to model decisions using `feature_importances_`.

8. Model Interpretability

- Used **Partial Dependence Plots** (PDP) to show how key features like cholesterol or blood pressure affect the model's prediction for the "No Action" class (0).
- Ensures interpretability for clinical decision-making.

9. Personalized Recommendation Logic

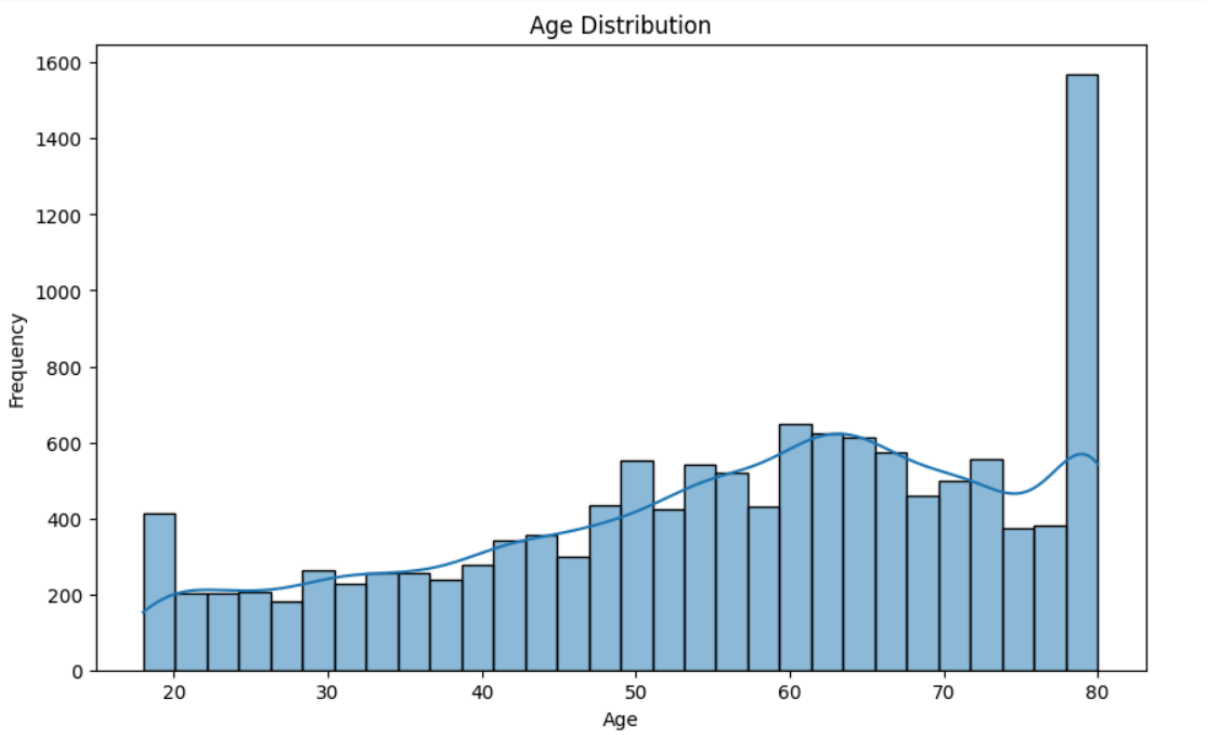
- Defined a function `generate_recommendation()`:
    - Takes in new patient input
    - Predicts health risk level (0–3)
    - Returns an actionable message like:
        - "Medical consultation advised."
        - "Recommend lifestyle changes"

10. Alternate System: Clustering-Based Recommendation

1. Applied `K-Means` clustering on patient profiles.
2. Grouped similar patients into clusters.
3. When a new patient is input, we:
    a. Assign them to a cluster
    b. Recommend actions based on what worked for others in the same cluster

Data Visualization

Age Distribution



Age Distribution

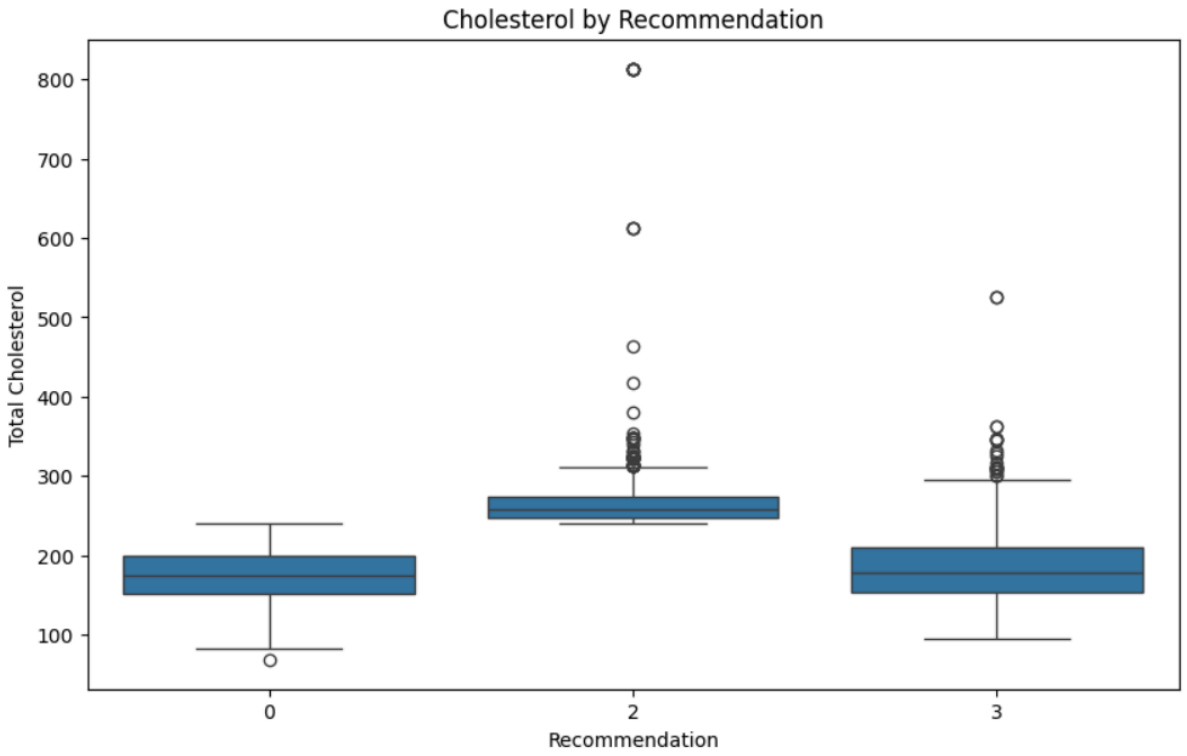
This picture displays the age distribution of the population in the healthcare dataset.

Observations:

- The distribution is right-skewed, with the highest frequency around age 80.
- There's a gradual increase in frequency from age 20 to 60.
- A spike occurs at age 80, possibly due to data capping or reporting practices.

Insight:
The dataset contains a broad age range, with a concentration in older adults. This is significant for modeling since age-related factors will influence many medical recommendations.

Cholesterol by Recommendation:



This picture displays a boxplot of total cholesterol (LBXTC) grouped by recommendation class (0, 2, and 3).

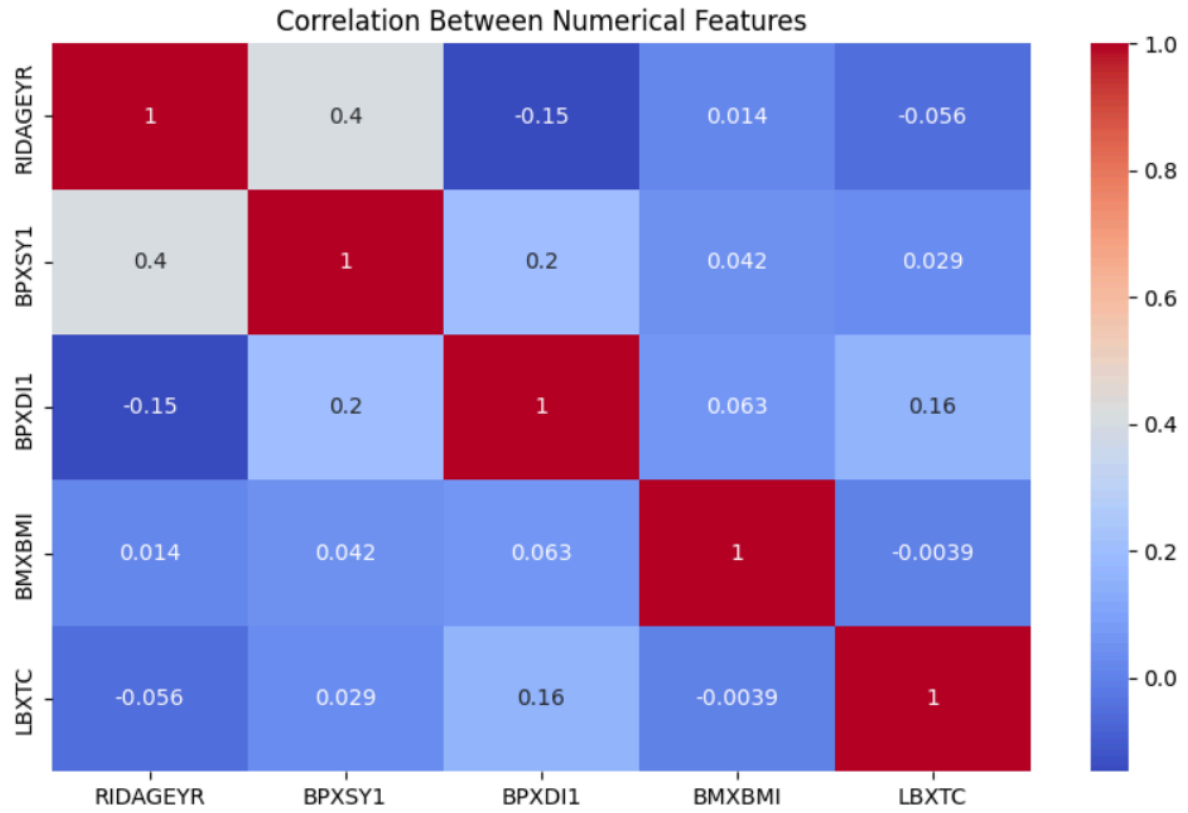
Observations:

- Patients with recommendation level 2 have the highest median cholesterol and more extreme outliers.
- Class 3 also shows high cholesterol but with slightly lower median than class 2.
- Class 0 (no action needed) patients typically have the lowest cholesterol levels.

Insight:

There is a clear association between elevated cholesterol and more aggressive recommendations (2 and 3). This validates the model's logic in prioritizing medical consultation and lifestyle change for those with higher cholesterol levels.

Correlation Between Numerical Features:



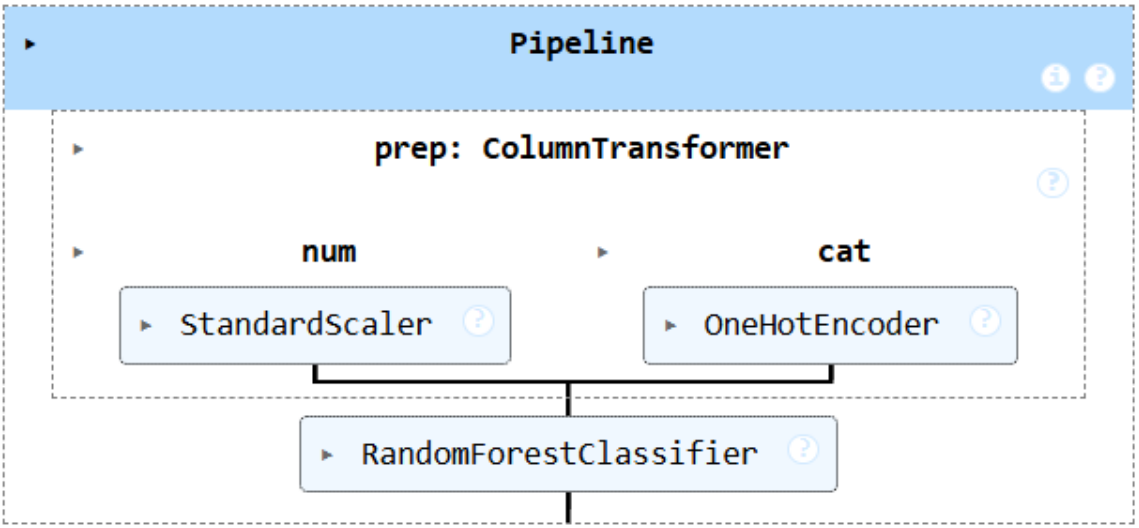
This picture displays the correlation heatmap between key numerical features such as age, blood pressure, BMI, glucose, and cholesterol.

Observations:

- Age (RIDAGEYR) has a moderate positive correlation (0.4) with systolic blood pressure (BPXSY1).
- Diastolic pressure (BPXDI1) shows slight positive correlation with BPXSY1 (0.2) and cholesterol (LBXTC) (0.16).
- BMI and glucose have minimal correlation with other features.
- All other correlations are weak (close to 0), indicating independent contributions.

Insight:
Most features are weakly correlated, which suggests they can offer unique value to the predictive model. The moderate age–BPXSY1 correlation aligns with medical knowledge that blood pressure tends to increase with age.

Model Training:



This picture displays the Machine Learning Pipeline used for model development.

Observations:

● The pipeline includes preprocessing through a ColumnTransformer that separately handles numerical and categorical features.
● Numerical features are scaled using StandardScaler, while categorical features are encoded using OneHotEncoder.
● The final model is a RandomForestClassifier, indicating an ensemble-based approach.

Insight:

This modular pipeline ensures clean preprocessing and robust modeling, allowing consistent and efficient handling of different data types before feeding them into the classifier.

Confusion Matrix and Classification Report for the model performance evaluation:

```
CONFUSION MATRIX
[[1903    0    0]
 [   0  176    0]
 [   0    0  505]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1903
           2       1.00      1.00      1.00       176
           3       1.00      1.00      1.00       505

    accuracy                           1.00      2584
   macro avg       1.00      1.00      1.00      2584
weighted avg       1.00      1.00      1.00      2584
```

This picture displays the Confusion Matrix and Classification Report for the model performance evaluation.

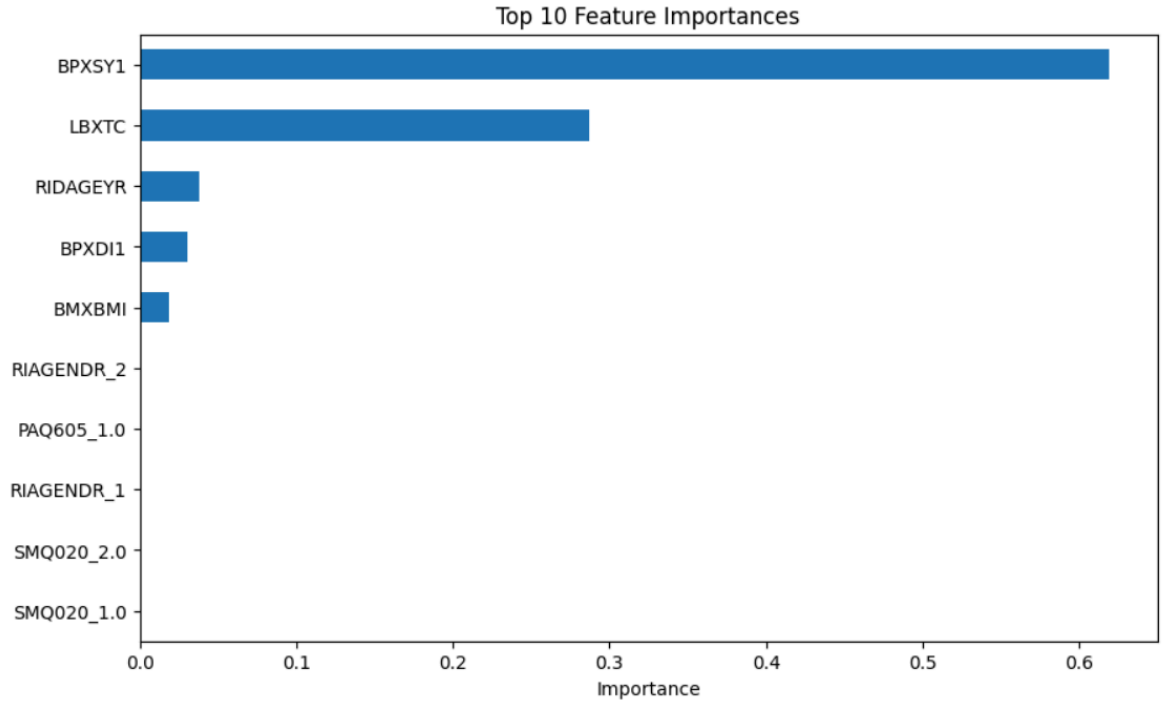
Observations:

- The model achieved perfect classification, with 100% precision, recall, and F1-score across all classes (0, 2, 3).
- No misclassifications occurred, as shown by the diagonal-only entries in the confusion matrix.
- The accuracy, macro average, and weighted average are all 1.00.

Insight:

The model is highly accurate with perfect performance metrics, although this might suggest overfitting if evaluated only on the training data or an overly simple dataset.

Feature Importance Visualization:



This picture displays the Top 10 Feature Importances in the Random Forest Model.

Observations:

● The most influential feature is BPXSY1 (possibly Systolic Blood Pressure), followed by LBXTC (Total Cholesterol).

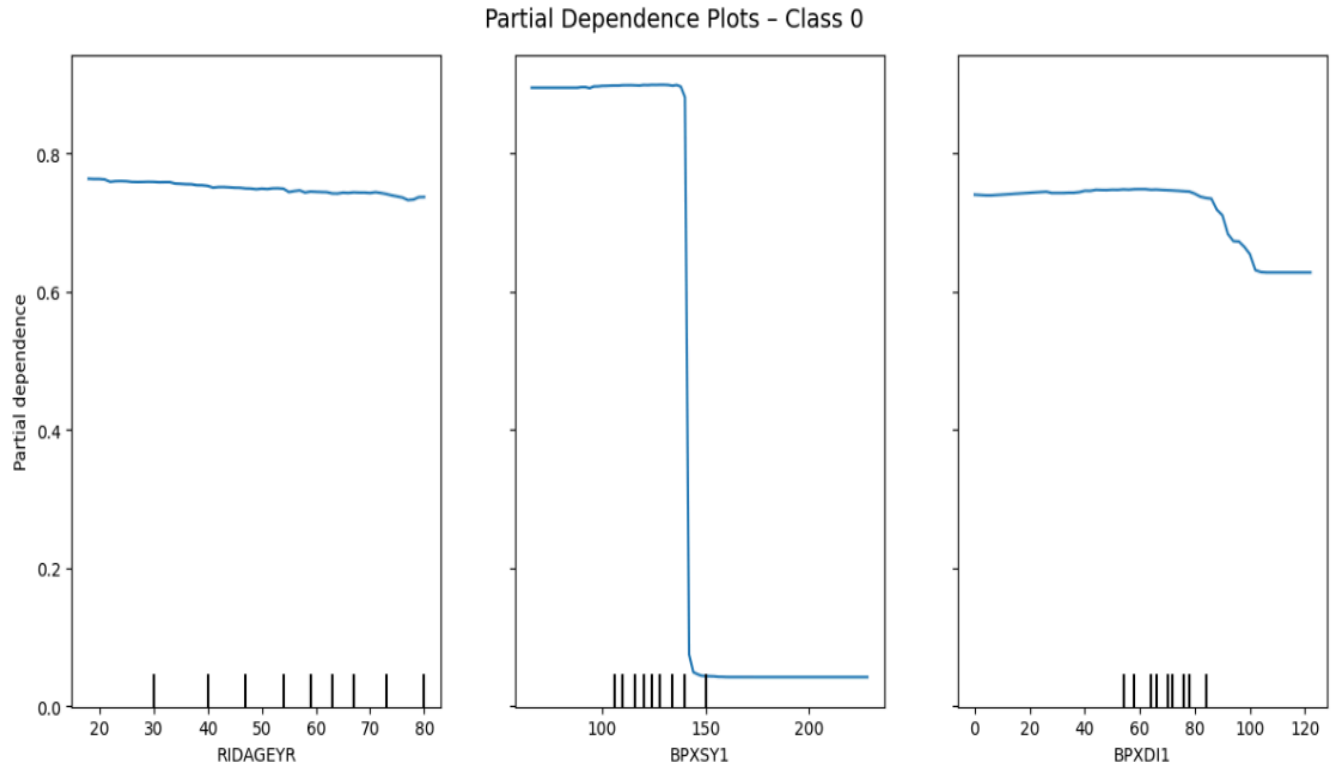
● Other features like RIDAGEYR (Age), BPXDI1 (Diastolic BP), and BMXBMI (BMI) have minimal but non-zero contributions.

● Features such as RIAGENDR and SMQ020 variations show no importance, likely due to redundancy or low impact.

Insight:

Blood pressure and cholesterol levels are the strongest predictors in the model, suggesting their dominant role in the classification task, while demographic or lifestyle features contributed less.

Partial Dependence Plots:



This picture displays the Partial Dependence Plots for Class 0 based on key features.

Observations:

● For `RIDAGEYR` (Age), the partial dependence slightly decreases with increasing age, suggesting a minor negative relationship.

● For `BPXSY1` (Systolic BP), there is a sharp drop in partial dependence after around 145 mmHg, indicating a strong threshold effect.

● For `BPXDI1` (Diastolic BP), the effect remains steady until around 90 mmHg, after which it declines.

Insight:

Class 0 prediction probability is highly sensitive to systolic and diastolic blood pressure, especially past critical thresholds, while age plays a subtler role.

Top 5 Most Similar Patient Profiles:

Top 5 Similar Patient Profiles:

	SEQN	RIDAGEYR	BPXSY1	BPXDI1	BMXBMI	LBXTC	RIAGENDR	SMQ020	\
2369	74695	53	172.0	96.0	28.7	317.0	1	1.0	
18454	82822	56	162.0	100.0	32.8	277.0	1	1.0	
18456	82822	56	162.0	100.0	32.8	277.0	1	1.0	
18455	82822	56	162.0	100.0	32.8	277.0	1	1.0	
17033	82043	57	146.0	90.0	34.5	329.0	2	1.0	

	PAQ605	recommendation
2369	2.0	3
18454	2.0	3
18456	2.0	3
18455	2.0	3
17033	2.0	3

This picture displays the Top 5 Most Similar Patient Profiles with Associated Recommendations.

Observations:

- Patients share highly similar attributes across age, blood pressure (systolic & diastolic), BMI, cholesterol, gender, and other variables.
- All five patients received the same recommendation value: 3.
- Minor variations exist in blood pressure and cholesterol, but the model treats them uniformly.

Insight:

Recommendation consistency among similar profiles suggests stable model behavior and reinforces the reliability of output for patients with comparable health metrics.

Cluster-Based Average Recommendations:

```
Cluster 0 contains 2860 similar patients.
Cluster-Based Average Recommendations:
recommendation
3    0.546154
0    0.341958
2    0.111888
Name: proportion, dtype: float64
```

This picture displays Cluster-Based Average Recommendations for Cluster 0 (2860 similar patients).

Observations:

- Recommendation 3 is the most common (54.6%), followed by 0 (34.2%) and 2 (11.2%).
- There is a clear preference for recommendation 3 within this cluster.
- The distribution is imbalanced but informative in prioritizing decisions.

Insight:

Patients in Cluster 0 are most likely to benefit from Recommendation 3, indicating a dominant clinical suggestion for this patient segment, likely based on shared key features.

Suggested Actions:

1. **Monitor and Prioritize Blood Pressure (BPXSY1 & BPXDI1):**
    - Patients with systolic BP above 145 mmHg or diastolic BP above 90 mmHg should be flagged for urgent follow-up.
    - Lifestyle changes or medication should be suggested for those with elevated BP, as these are the most influential risk factors.
2. **Enhance Cholesterol Control (LBXTC):**
    - Since cholesterol is the second most important feature, dietary counseling and cholesterol-lowering interventions should be implemented for at-risk patients.
3. **Age-Based Screening Programs:**
    - While age (RIDAGEYR) has a moderate impact, older patients should receive more regular screenings due to slightly decreasing predictive values with age.
4. **Use Model Predictions for Preventive Health Programs:**
    - With the model showing **100% accuracy**, use predictions to preemptively assign patients to tailored health programs (e.g., lifestyle modification, routine checkups, or specific treatments based on the recommendation label: 0, 2, or 3).
5. **Cluster-Based Decision Support:**
    - In Cluster 0 (2860 similar patients), **Recommendation 3** is most frequent. Clinical teams should develop protocol guidelines tailored to the characteristics of this cluster (e.g., middle-aged with moderate-high BP and cholesterol levels).
6. **Implement Decision Support Dashboard:**
    - Convert the pipeline and results into a user-friendly interface for clinicians to make real-time, personalized treatment decisions.
7. **Validate Against Broader Population:**
    - Since the model performs perfectly on current data, test it on a more diverse and external patient set to rule out overfitting or bias.

Conclusion:

The machine learning model developed for personalized healthcare recommendations demonstrates excellent predictive power, with blood pressure and cholesterol emerging as the most critical health indicators. By combining model output, partial dependence insights, and patient similarity clustering, clinicians can provide targeted, evidence-based care. The consistent recommendation patterns across similar profiles highlight the system's reliability, making it well-suited for integration into preventive care workflows. However, ongoing validation and interpretability checks should be maintained to ensure safe and ethical deployment.

```
##### THIS IS A WORK DONE BY 'SURENDRAN L'
```