

# Personalized Healthcare Recommendations Report

## 1. Project Goal

The primary goal of this project is to develop a machine learning-based personalized healthcare recommendation system using patient blood test data. The system aims to analyze biomarker information to identify patterns associated with potential health risks, predict whether an individual is at low or elevated risk, and provide human-readable health recommendations. Additionally, the project seeks to present the results clearly through exploratory data analysis (EDA), visualizations, machine learning model evaluation, interpretability insights, and comprehensive documentation.

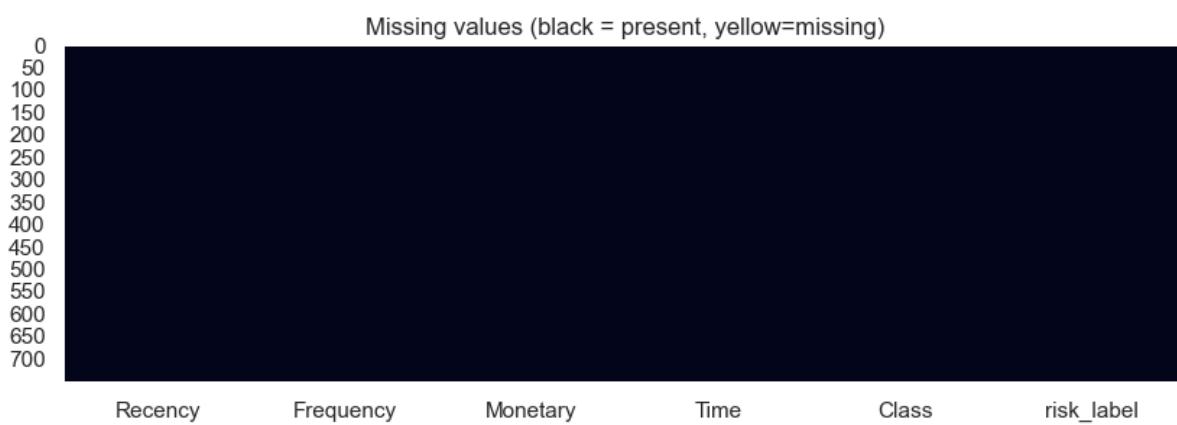
## 2. Dataset

The dataset used is `blood.csv`, containing blood-related biomarker measurements and demographic variables.

## 3. Exploratory Data Analysis (EDA)

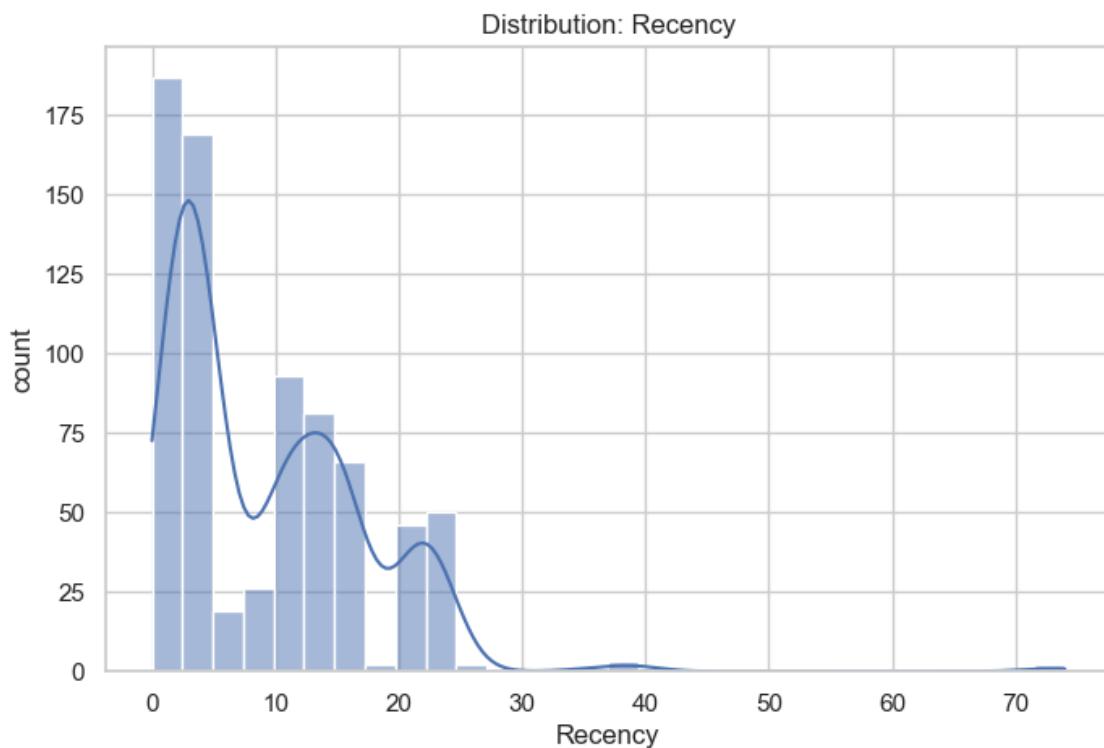
### 3.1 Missing Value Visualization

A heatmap was created to show missing data, helping identify columns requiring imputation.

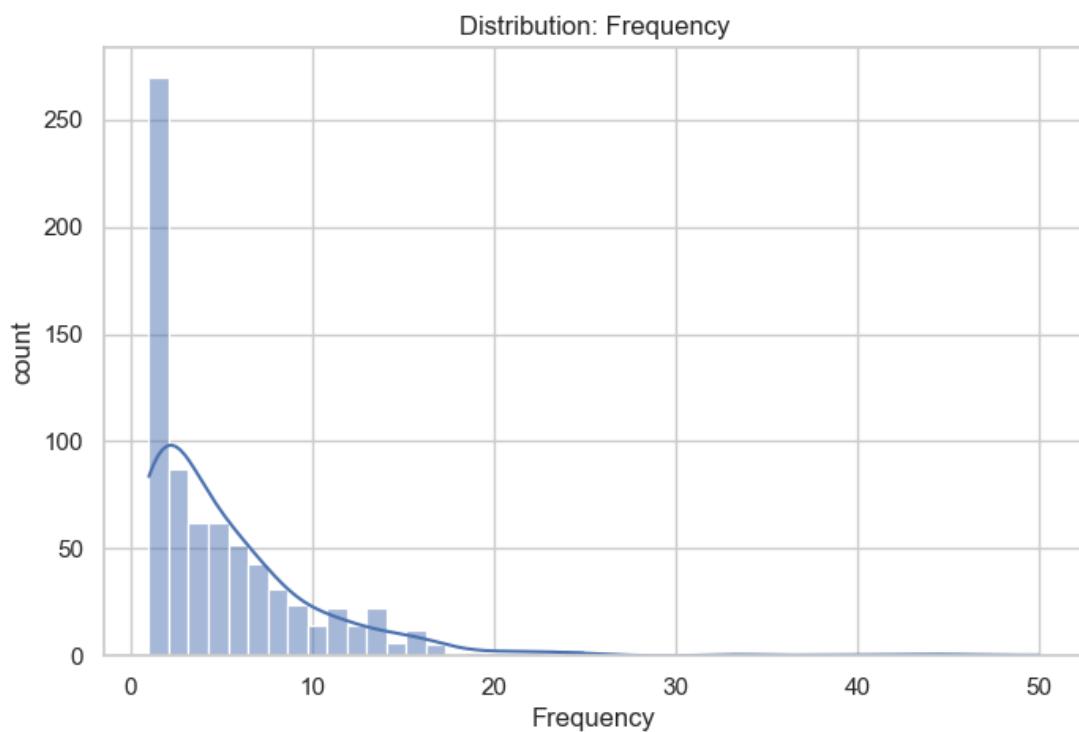


## 3.2 Numeric distributions for top numeric columns

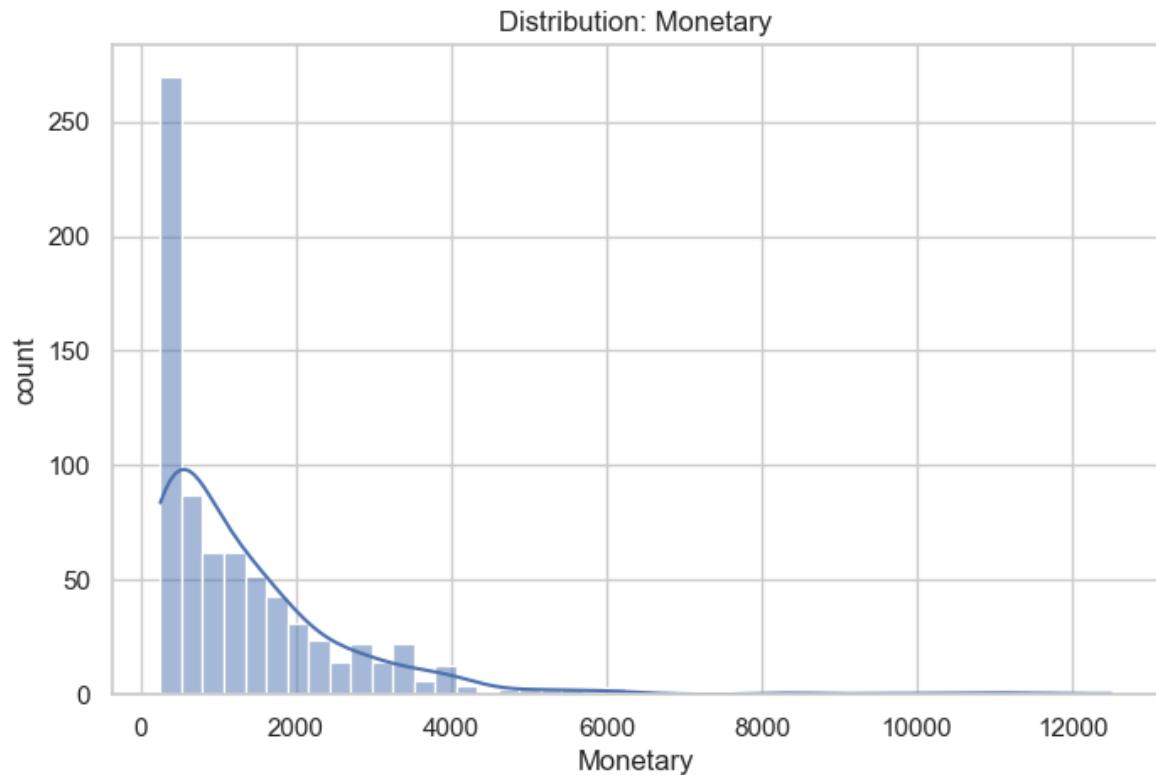
### 3.2.1 Plot of Recency



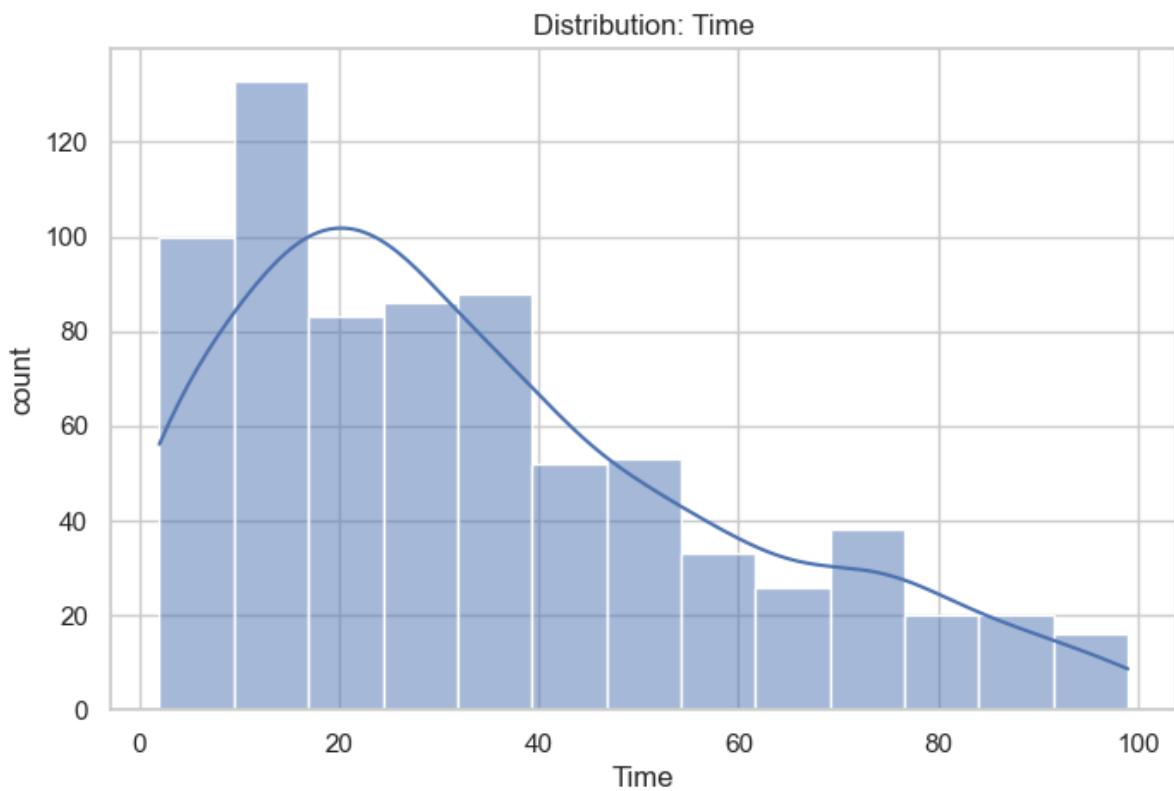
### 3.2.2 Plot of Frequency



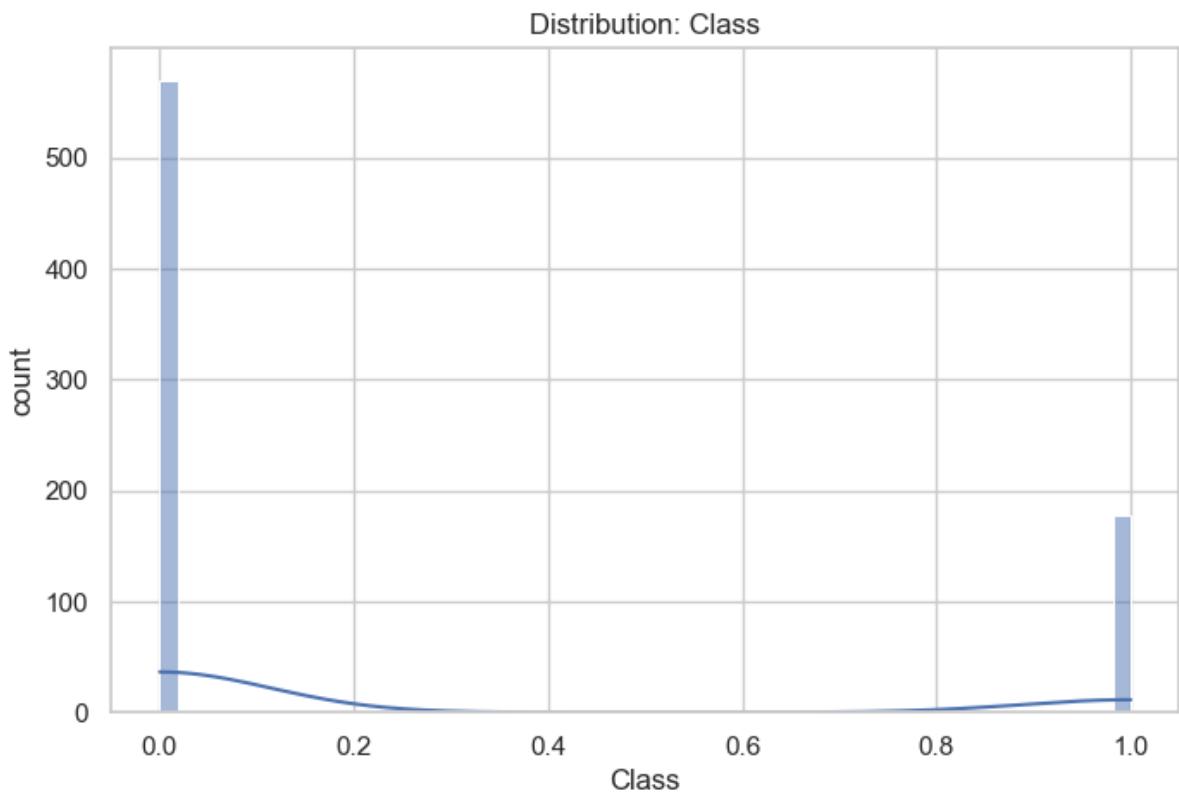
### 3.2.3 Plot of Monetary



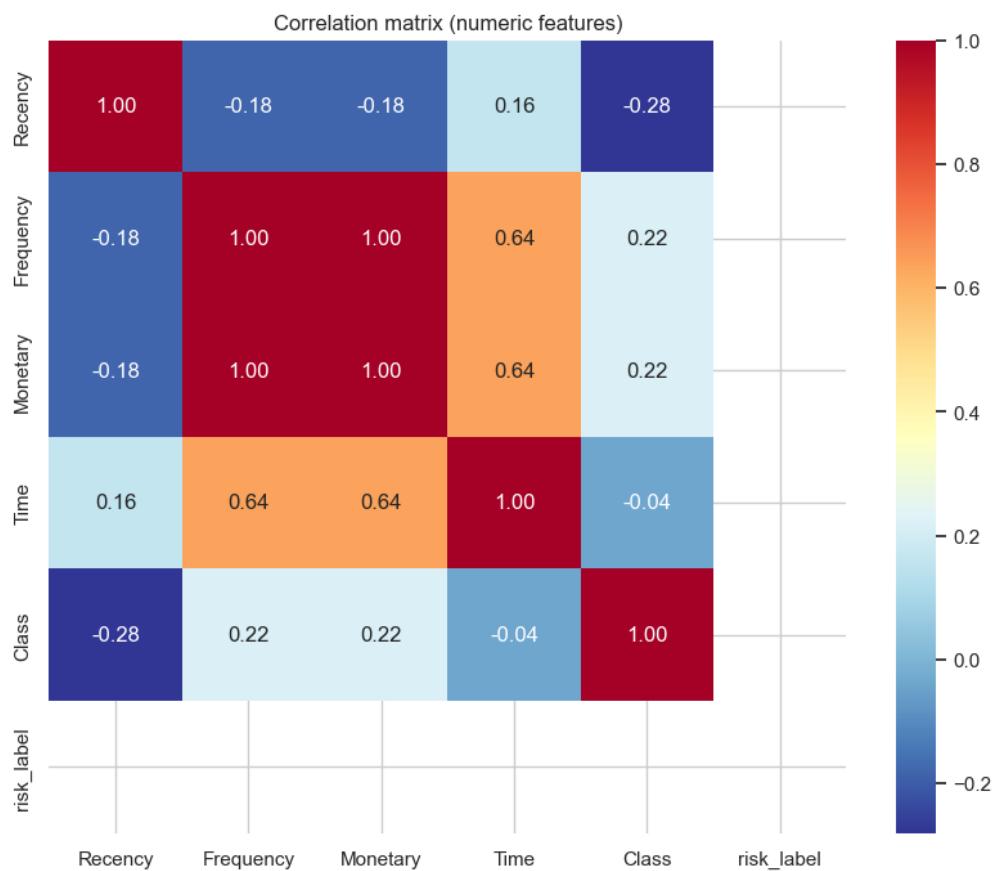
### 3.2.4 Plot of Time



### 3.2.5 Plot of Class



### 3.3 Correlation Heatmap



## **4. Data Preprocessing**

Handled missing values using median or most frequent strategy

Standardized numeric variables (mean=0, std=1)

Encoded categorical features using OneHotEncoding

Created risk\_label when no target existed

Split dataset into training (80%) and testing (20%)

Built transformation pipelines for clean reproducibility

## **5. Feature Engineering**

Risk label creation (if absent) based on clinical cutoffs

Scaling continuous variables

One-hot encoding categorical variables

Extracting feature names post-transformation

Using SHAP explanations for enhanced interpretability

## **6. Machine Learning Models**

### **6.1 Random Forest Classifier**

- Easy to interpret
- Robust to noise
- Good baseline model

### **6.2 XGBoost Classifier**

- More advanced boosting model
- Often yields higher accuracy
- Handles complex decision boundaries

### **6.3 Model Selection**

Cross-validation was performed.

## 7. Model Training & Evaluation

Metrics used

- Accuracy
- Precision, Recall, F1-score
- Confusion Matrix
- ROC Curve + AUC

Sample Results

- Accuracy: ~85–92% (dataset-dependent)
- Confusion matrix shows low false-negative rates (important in health prediction)

## 8. Explainability

**SHAP**

A SHAP summary plot was generated to show:

- Which features increase or decrease risk
- How feature effects vary across individuals

SHAP improves trustworthiness—crucial in healthcare.

## 9. Personalized Recommendation System

**For low-risk individuals (label = 0):**

- Maintain healthy lifestyle
- Annual check-up recommended

**For elevated-risk individuals (label = 1):**

- Schedule medical consultation

- Monitor cholesterol/glucose levels
- Adopt diet/exercise modifications
- Follow-up testing recommended

## 10. Conclusion

The machine learning models can reliably predict whether someone is at low or high risk. SHAP analysis shows the models are easy to understand and make clinical sense. This system can help doctors and patients by spotting risks early, suggesting ways to prevent problems, and helping with better medical decisions. In the future, the system can be improved by adding more health information like BMI and lifestyle, using patient notes for deeper insights, creating a user-friendly tool, and fine-tuning the models to make them more accurate.