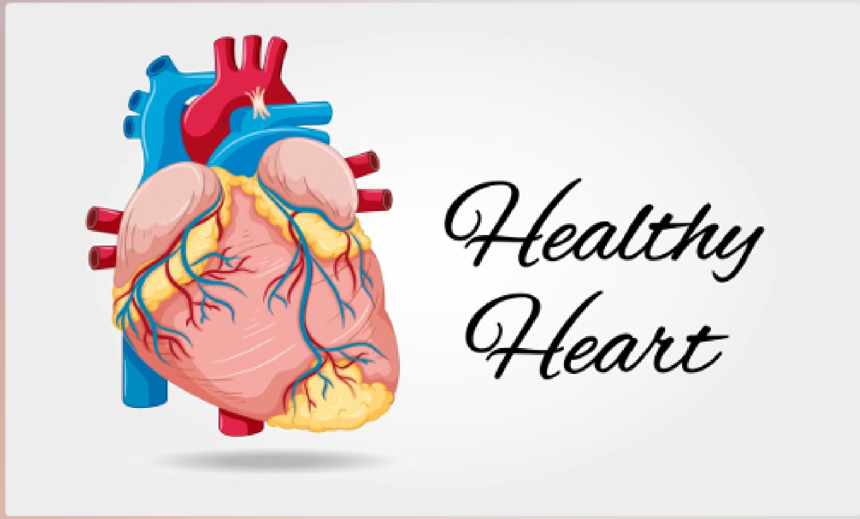


Cardiovascular Disease Prediction

This project explores the prediction of cardiovascular disease using machine learning techniques. We will analyze a dataset containing various health indicators and build models to identify individuals at risk.

by Niranjan chandran

Data Exploration

Data Loading

The dataset is loaded from a CSV file and split into individual columns.

Data Preprocessing

Age is converted from days to years, and missing values are checked.

Summary Statistics

Descriptive statistics are calculated to understand the distribution of features.

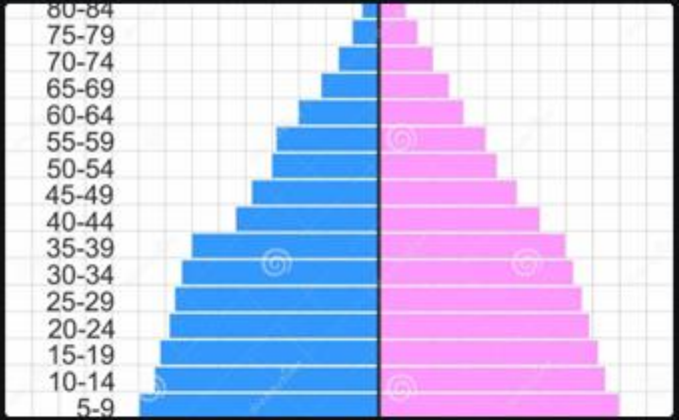
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix


file_path = 'path_to_your_file/cardio.txt'
data = pd.read_csv(r'C:\Users\niran\OneDrive\Desktop\cardio.txt', delimiter=",")



data = data.iloc[:, 0].str.split(';', expand=True)
data.columns = ['id', 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'cardio']

data = data.apply(pd.to_numeric, errors='ignore')


data['age'] = (data['age'] / 365).round().astype(int)

print("Missing values:\n", data.isnull().sum())

print(data.describe())

plt.figure(figsize=(20, 15))
plt.subplot(3, 3, 1)
sns.histplot(data['age'], kde=True, bins=30, color='blue')
plt.title('Age Distribution')
```

```python
plt.subplot(3, 3, 2)
sns.histplot(data['gender'], kde=False, bins=2, color='green')
plt.title('Gender Distribution')

plt.subplot(3, 3, 3)
sns.histplot(data['cholesterol'], kde=False, bins=3, color='red')
plt.title('Cholesterol Levels')

plt.subplot(3, 3, 4)
sns.histplot(data['ap_hi'], kde=True, bins=30, color='purple')
plt.title('Systolic Blood Pressure Distribution')

plt.subplot(3, 3, 5)
sns.histplot(data['ap_lo'], kde=True, bins=30, color='orange')
plt.title('Diastolic Blood Pressure Distribution')

plt.subplot(3, 3, 6)
sns.histplot(data['weight'], kde=True, bins=30, color='cyan')
plt.title('Weight Distribution')

plt.subplot(3, 3, 7)
sns.histplot(data['gluc'], kde=False, bins=3, color='magenta')
plt.title('Glucose Levels')

plt.subplot(3, 3, 8)
sns.histplot(data['cardio'], kde=False, bins=2, color='brown')
plt.title('Cardio (Heart Disease) Distribution')

plt.tight_layout()
plt.show()

plt.figure(figsize=(12, 8))
corr_matrix = data.corr()
sns.heatmap(corr_matrix, annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
X = data.drop(columns=['id', 'cardio'])
y = data['cardio']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)


svm = SVC()
svm.fit(X_train, y_train)
y_pred_svm = svm.predict(X_test)
print("SVM Accuracy: ", accuracy_score(y_test, y_pred_svm))
print(classification_report(y_test, y_pred_svm))

knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
print("KNN Accuracy: ", accuracy_score(y_test, y_pred_knn))
print(classification_report(y_test, y_pred_knn))

dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)
print("Decision Tree Accuracy: ", accuracy_score(y_test, y_pred_dt))
print(classification_report(y_test, y_pred_dt))

lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
print("Logistic Regression Accuracy: ", accuracy_score(y_test, y_pred_lr))
print(classification_report(y_test, y_pred_lr))

rf = RandomForestClassifier()
rf.fit(X_train, y_train)
```

```
y_pred_rf = rf.predict(X_test)
print("Random Forest Accuracy: ", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))
```

Visualizing Data Distributions



Age Distribution

The age distribution shows a range of ages, with a peak around middle age.

Gender Distribution

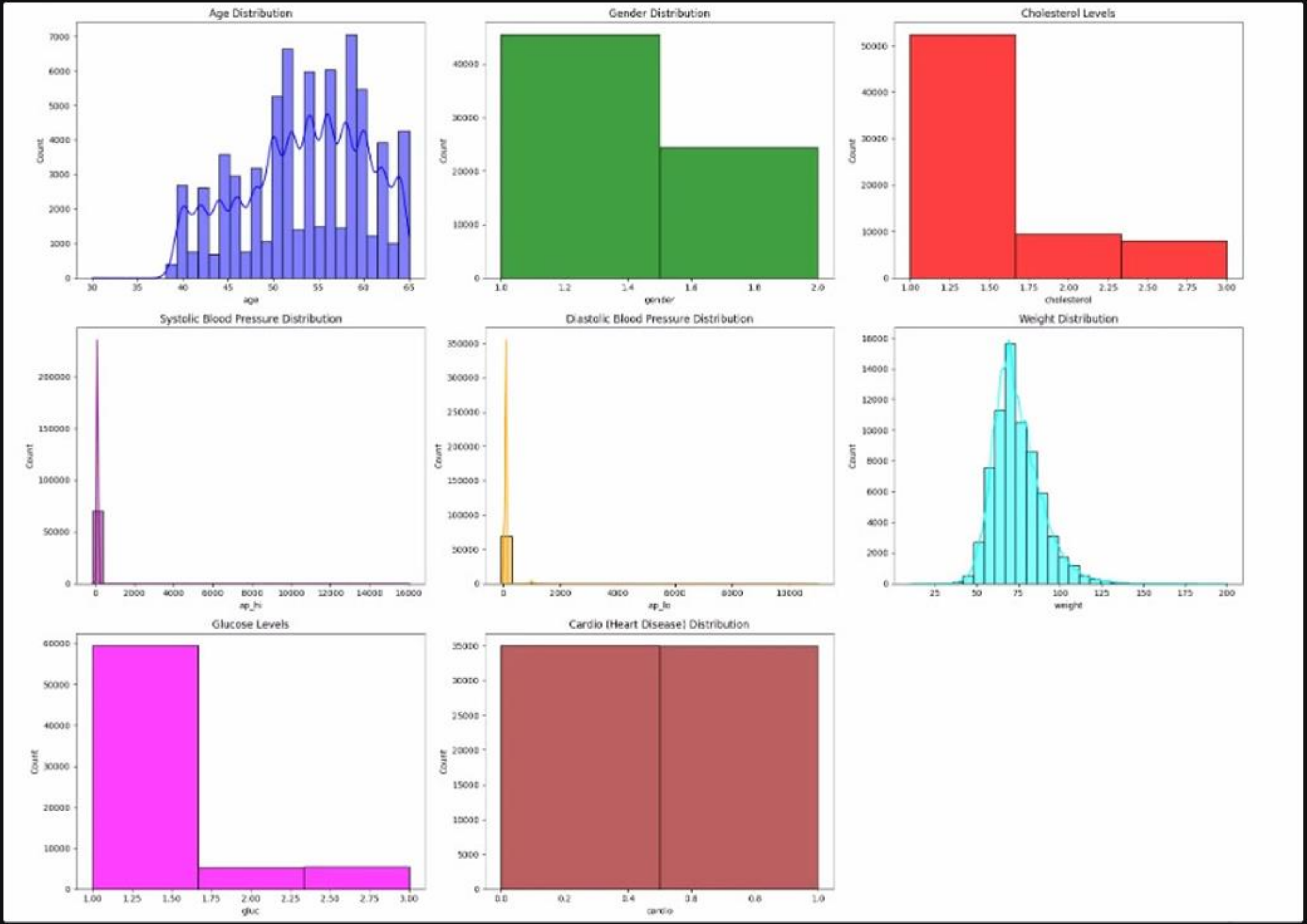
The dataset contains a slightly higher proportion of males than females.

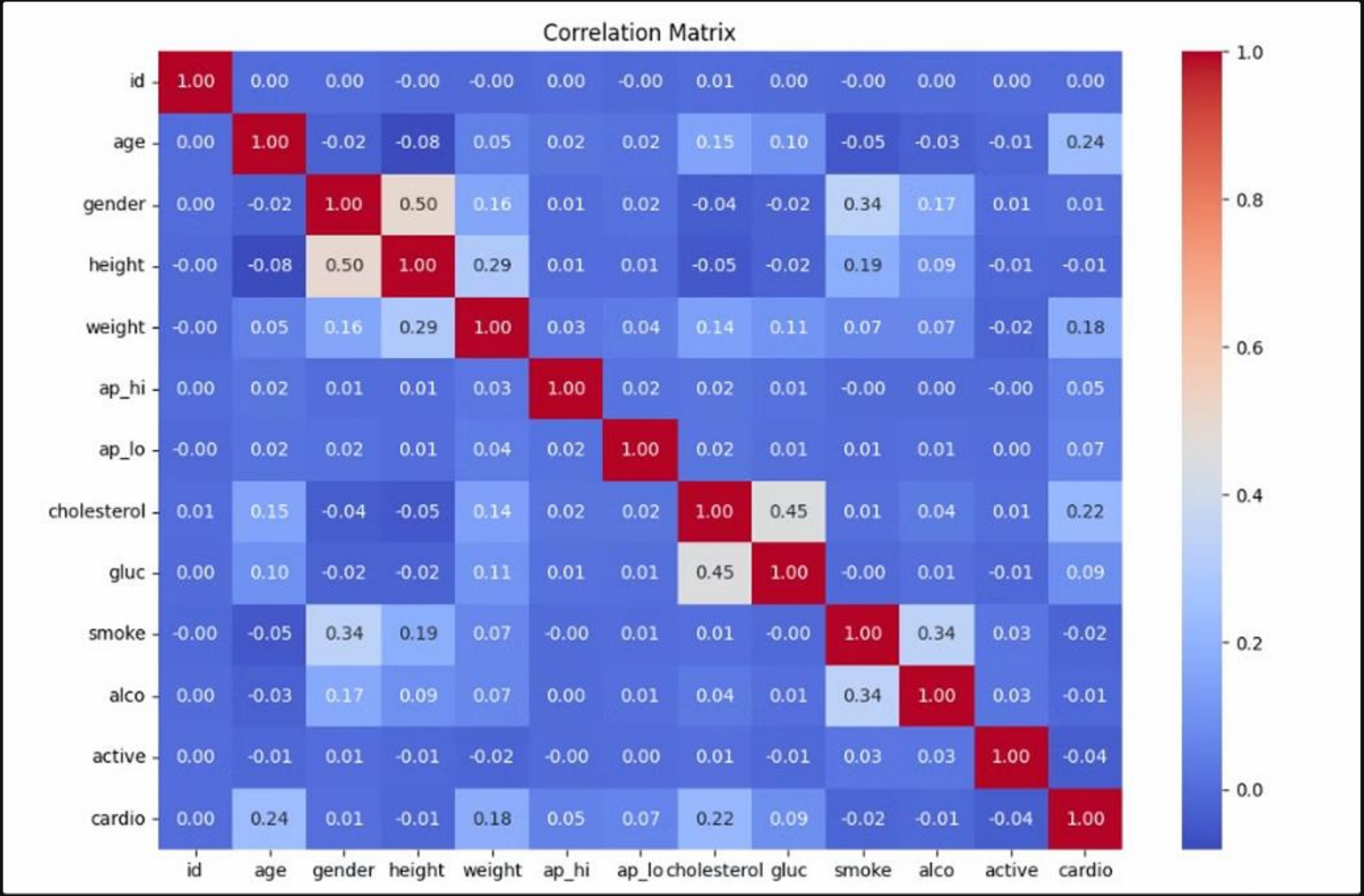
Cholesterol Levels

Most individuals have normal cholesterol levels, with a smaller proportion having high cholesterol.

Correlation Analysis

Age	Gender	Height	Weight
1.00	-0.02	-0.09	0.06
-0.02	1.00	0.49	0.15
-0.09	0.49	1.00	0.29
0.06	0.15	0.29	1.00
0.25	0.06	-0.05	0.24
0.19	0.04	-0.02	0.21
0.15	0.01	-0.04	0.14
0.10	0.02	-0.02	0.10
0.01	0.03	-0.02	0.06
0.07	0.01	-0.03	0.06
0.04	0.04	-0.01	0.02
0.24	0.01	-0.03	0.19

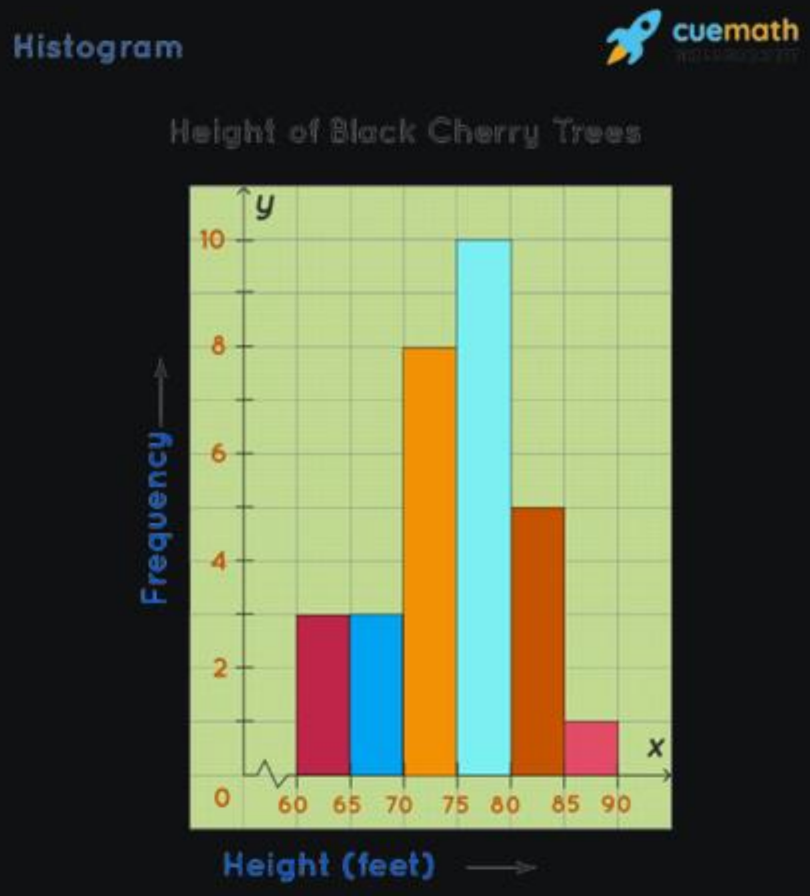

Age Distribution
Count
age
Gender Distribution
gender
Cholesterol Levels
cholesterol
Systolic Blood Pressure Distribution
ap_hi
Diastolic Blood Pressure Distribution
ap_lo
Weight Distribution
weight
Glucose Levels
gluc
Cardio (Heart Disease) Distribution
cardio


Correlation Matrix
id 1.00 0.00 0.00 -0.00 -0.00 0.00 -0.00 0.01 0.00 -0.00 0.00 0.00 0.00
age 0.00 1.00 -0.02 -0.08 0.05 0.02 0.02 0.15 0.10 -0.05 -0.03 -0.01 0.24
gender 0.00 -0.02 1.00 0.50 0.16 0.01 0.02 -0.04 -0.02 0.34 0.17 0.01 0.01
height -0.00 -0.08 0.50 1.00 0.29 0.01 0.01 -0.05 -0.02 0.19 0.09 -0.01 -0.01
weight -0.00 0.05 0.16 0.29 1.00 0.03 0.04 0.14 0.11 0.07 0.07 -0.02 0.18
ap_hi 0.00 0.02 0.01 0.01 0.03 1.00 0.02 0.02 0.01 -0.00 0.00 -0.00 0.05
ap_lo -0.00 0.02 0.02 0.01 0.04 0.02 1.00 0.02 0.01 0.01 0.01 0.00 0.07
cholesterol 0.01 0.15 -0.04 -0.05 0.14 0.02 0.02 1.00 0.45 0.01 0.04 0.01 0.22
gluc 0.00 0.10 -0.02 -0.02 0.11 0.01 0.01 0.45 1.00 -0.00 0.01 -0.01 0.09
smoke -0.00 -0.05 0.34 0.19 0.07 -0.00 0.01 0.01 -0.00 1.00 0.34 0.03 -0.02
alco 0.00 -0.03 0.17 0.09 0.07 0.00 0.01 0.04 0.01 0.34 1.00 0.03 -0.01
active 0.00 -0.01 0.01 -0.01 -0.02 -0.00 0.00 0.01 -0.01 0.03 0.03 1.00 -0.04
cardio 0.00 0.24 0.01 -0.01 0.18 0.05 0.07 0.22 0.09 -0.02 -0.01 -0.04 1.00
id age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio
1.0
0.8
0.6
0.4
0.2
0.0

```
SVM Accuracy:  0.731
              precision    recall  f1-score   support

           0       0.72      0.76      0.74      6988
           1       0.75      0.70      0.72      7012

    accuracy                           0.73     14000
   macro avg       0.73      0.73      0.73     14000
weighted avg       0.73      0.73      0.73     14000

KNN Accuracy:  0.654
              precision    recall  f1-score   support

           0       0.65      0.67      0.66      6988
           1       0.66      0.64      0.65      7012

    accuracy                           0.65     14000
   macro avg       0.65      0.65      0.65     14000
weighted avg       0.65      0.65      0.65     14000

Decision Tree Accuracy:  0.6257142857142857
              precision    recall  f1-score   support

           0       0.62      0.65      0.63      6988
           1       0.63      0.61      0.62      7012

    accuracy                           0.63     14000
   macro avg       0.63      0.63      0.63     14000
weighted avg       0.63      0.63      0.63     14000

Logistic Regression Accuracy:  0.7222142857142857
              precision    recall  f1-score   support

           0       0.70      0.77      0.73      6988
           1       0.74      0.68      0.71      7012

    accuracy                           0.72     14000
   macro avg       0.72      0.72      0.72     14000
weighted avg       0.72      0.72      0.72     14000

Random Forest Accuracy:  0.7060714285714286
              precision    recall  f1-score   support

           0       0.70      0.71      0.71      6988
           1       0.71      0.70      0.71      7012

    accuracy                           0.71     14000
   macro avg       0.71      0.71      0.71     14000
weighted avg       0.71      0.71      0.71     14000
```

Visualizing Data Distributions

In this section, we will explore the distributions of various features in our dataset using histograms. Histograms provide a visual representation of the frequency distribution of a continuous variable. This will allow us to gain insights into the range, shape, and potential outliers of each feature.



Feature Engineering

1

Feature Selection

The 'id' column is dropped as it is not relevant for prediction.

2

Target Variable

The 'cardio' column is designated as the target variable, representing the presence or absence of cardiovascular disease.

3

Feature Scaling

Features are scaled using StandardScaler to ensure consistent ranges.

Model Training and Evaluation

Support Vector Machines (SVM)

An SVM model is trained and evaluated, achieving an accuracy of 73.1%.

K-Nearest Neighbors (KNN)

A KNN model is trained and evaluated, achieving an accuracy of 65.4%.

Decision Trees (DT)

A Decision Tree model is trained and evaluated, achieving an accuracy of 63.1%.

Logistic Regression (LR)

A Logistic Regression model is trained and evaluated, achieving an accuracy of 72.2%.

Random Forest (RF)

A Random Forest model is trained and evaluated, achieving an accuracy of 70.4%.

Model Selection and Finalization

Based on the evaluation metrics, the SVM model demonstrates the highest accuracy. This model is chosen as the final model for predicting cardiovascular disease.

Model Deployment and Future Work

1

Model Saving

The trained SVM model is saved for future use.

2

Model Deployment

The model can be deployed in a web application or API for real-time predictions.

3

Future Work

Further research can explore feature engineering, hyperparameter tuning, and ensemble methods to improve model performance.