Cardiovascular Health Assessment and Risk Prediction

1

Introduction

2

EDA Of Data

3

Feature Engineering

4

Discussion and Conclusion

Cardiovascular Health Assessment and Risk Prediction

Introduction

The goal is to predict the 10-year risk of Coronary Heart Disease (CHD) for patients based on demographic, behavioral, and medical risk factors. This involves developing a predictive model using a dataset of 3390 records and 16 attributes. The aim is to identify individuals at higher risk of CHD to enable early intervention and prevention strategies.

EDA Of Data

Feature Engineering

Discussion and Conclusion

Cardiovascular Health Assessment and Risk Prediction

Introduction

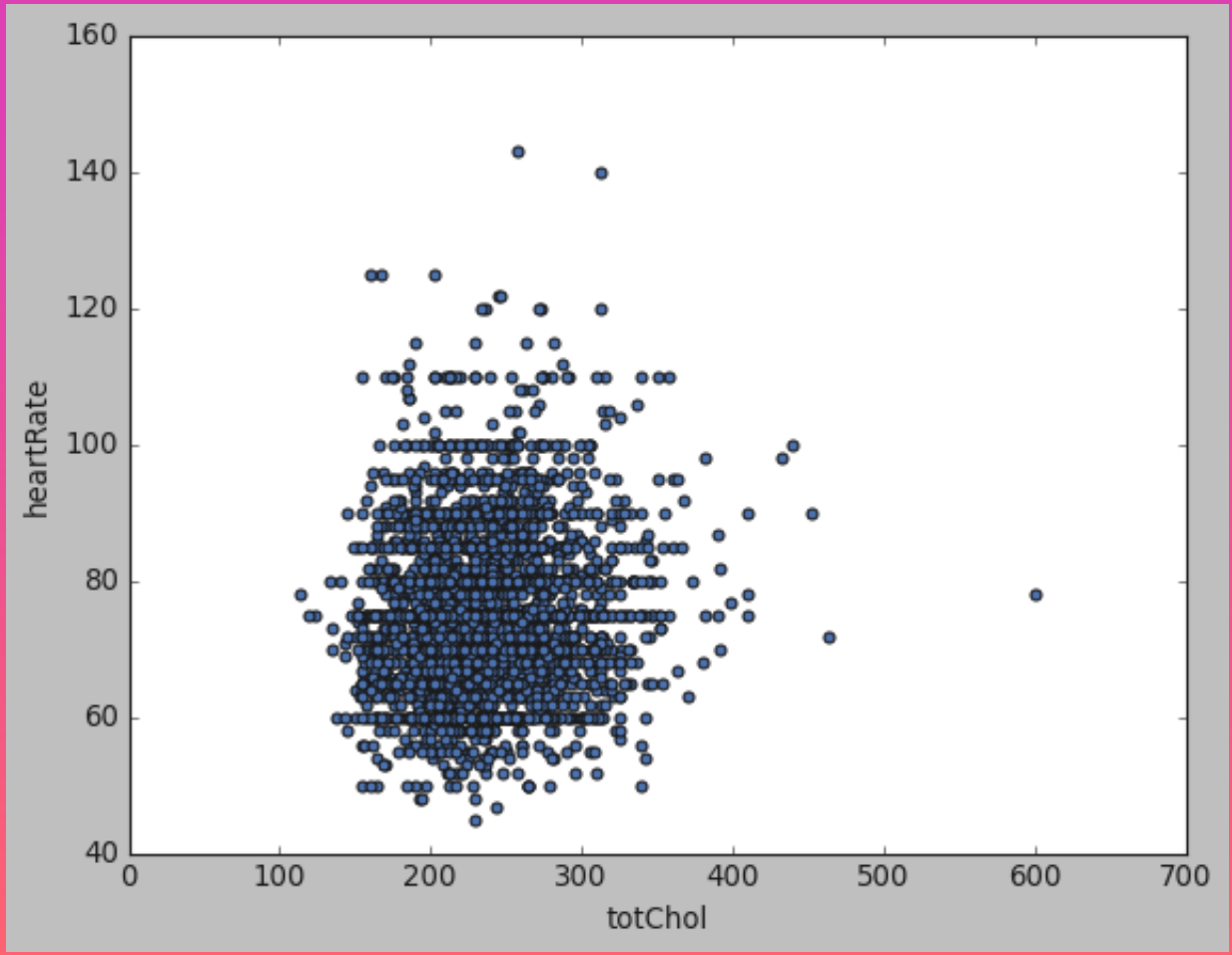
EDA

The succinct EDA involves understanding dataset structure, handling missing values, and identifying data types. Univariate analysis visualizes numerical and categorical variable distributions. Bivariate analysis explores relationships, especially with the target variable (10-year risk of CHD). Data visualization aids in comparison and correlation visualization.

Feature Engineering

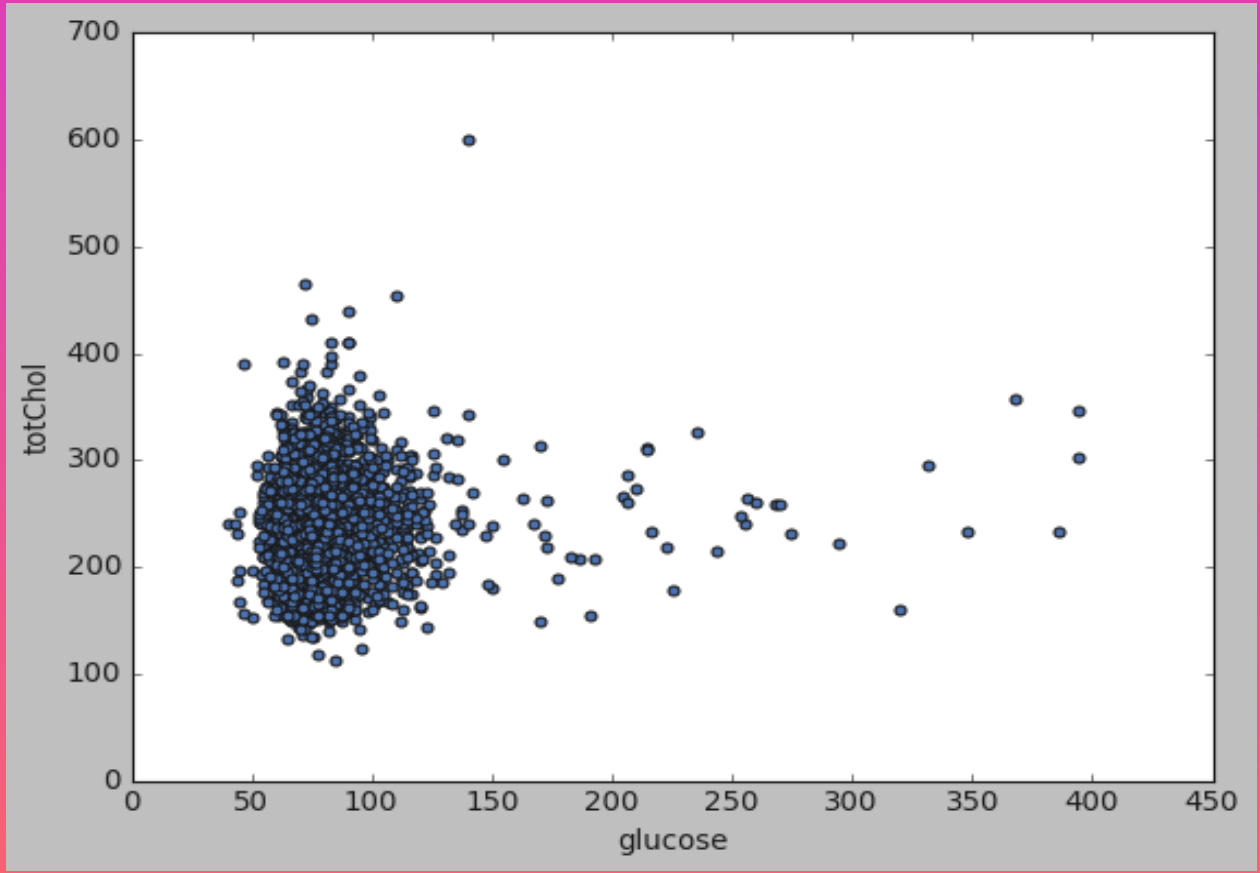
Discussion and Conclusion

Cardiovascular Health Assessment and Risk Prediction



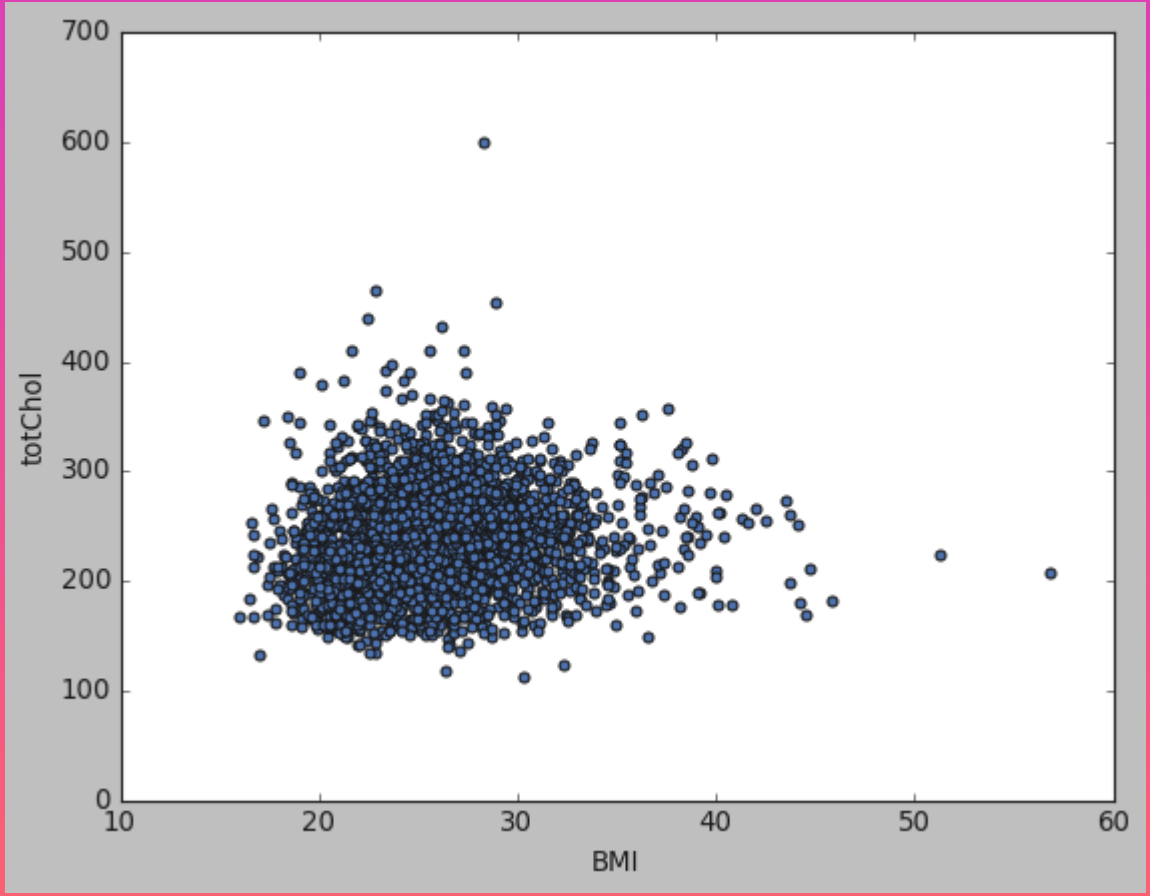
The majority of individuals seem to exhibit heart rates predominantly between 40 and 110 beats per minute and total cholesterol levels primarily between 120 and 450 mg/dL. This distribution implies variability in both heart rate and total cholesterol levels among the observed population. Further analysis could investigate potential correlations between these variables or explore any patterns or outliers within specific subgroups.

Cardiovascular Health Assessment and Risk Prediction



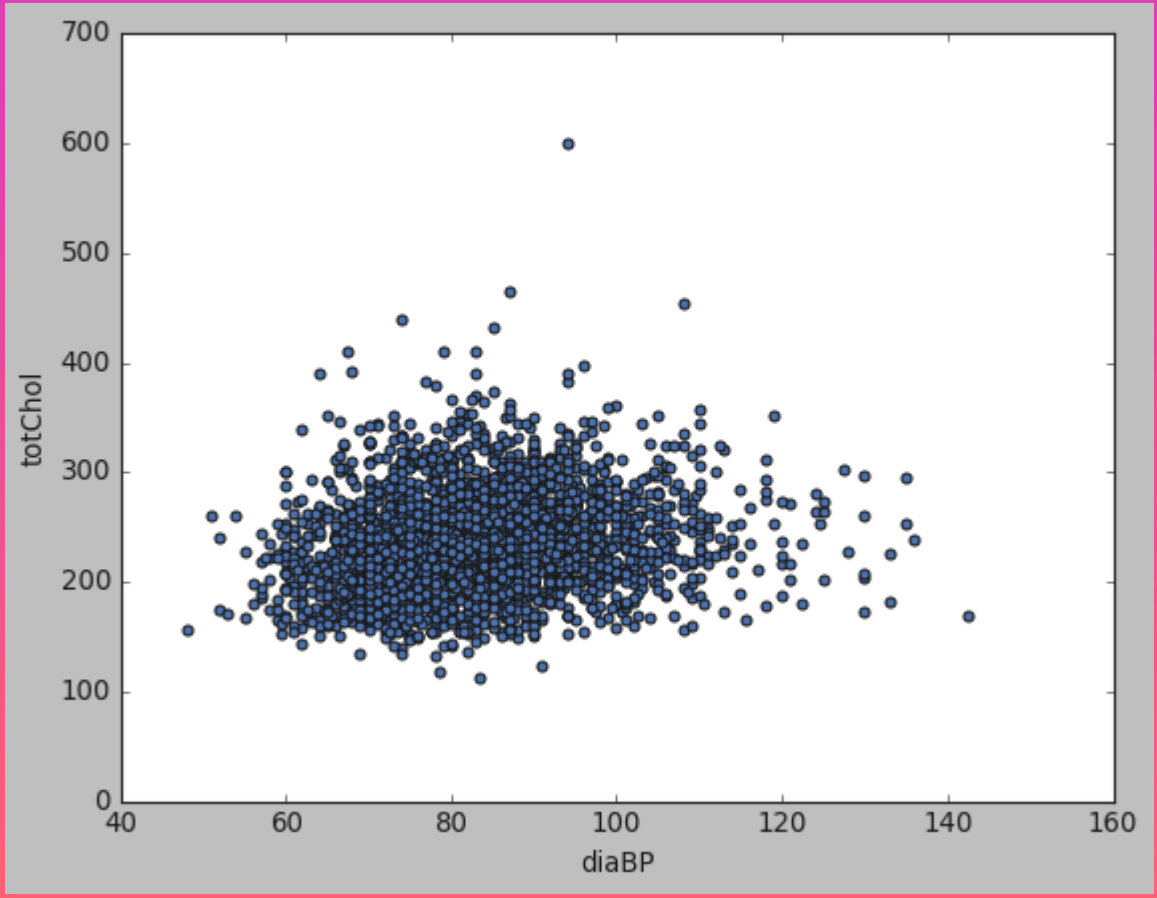
It appears there is a positive correlation between total cholesterol and glucose levels. However, there seems to be a concentration of data points around the lower end of both variables, with fewer data points at higher values. This concentration may indicate a common range of cholesterol and glucose levels among the observed population

Cardiovascular Health Assessment and Risk Prediction



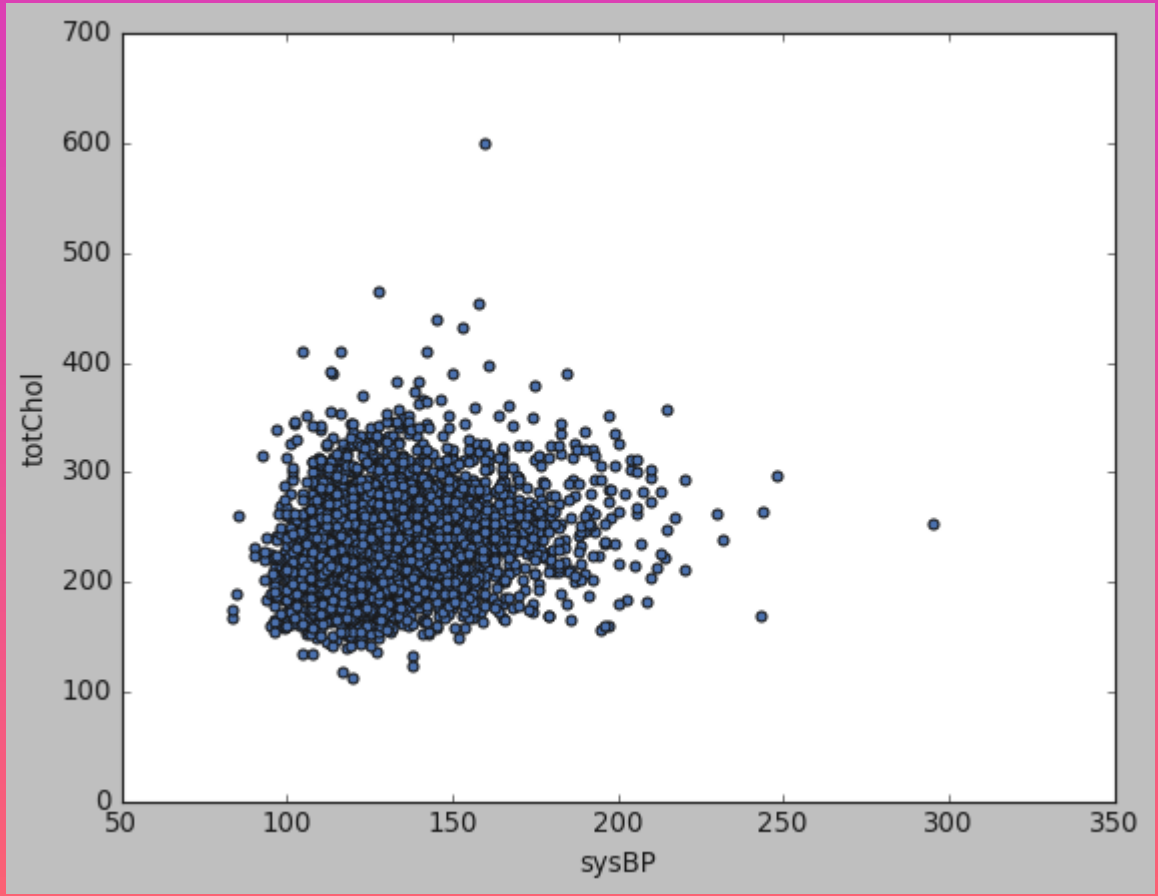
the scatter plot indicates a dispersed distribution of data points. The majority of individuals seem to exhibit BMI values predominantly between 15 and 40 kg/m^2 and total cholesterol levels primarily between 120 and 450 mg/dL. This distribution implies variability in both BMI and total cholesterol levels among the observed population.

Cardiovascular Health Assessment and Risk Prediction



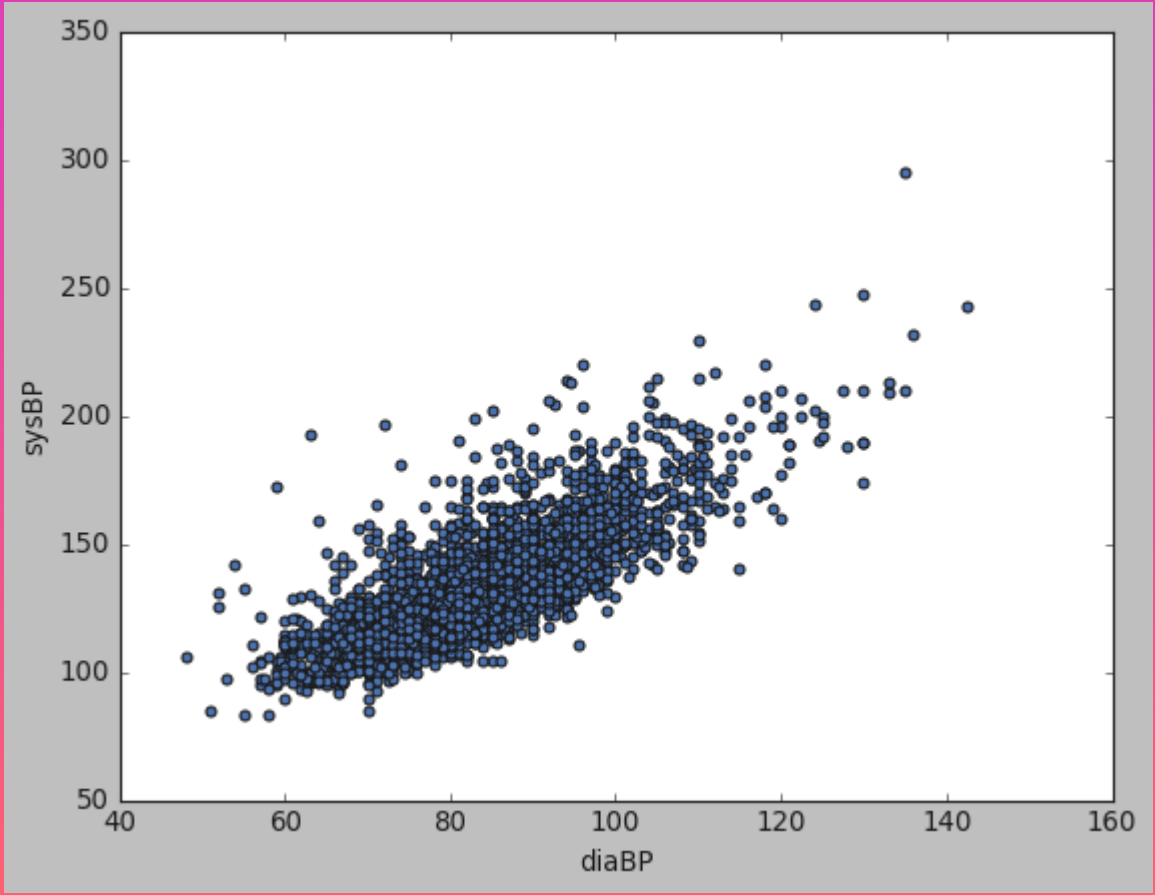
we have diastolic blood pressure (DiaBP) values, ranging from 40 to 160. Most dots are clustered between 120 and 450 total cholesterol and between 50 and 110 DiaBP, showing where most people's values fall.

Cardiovascular Health Assessment and Risk Prediction



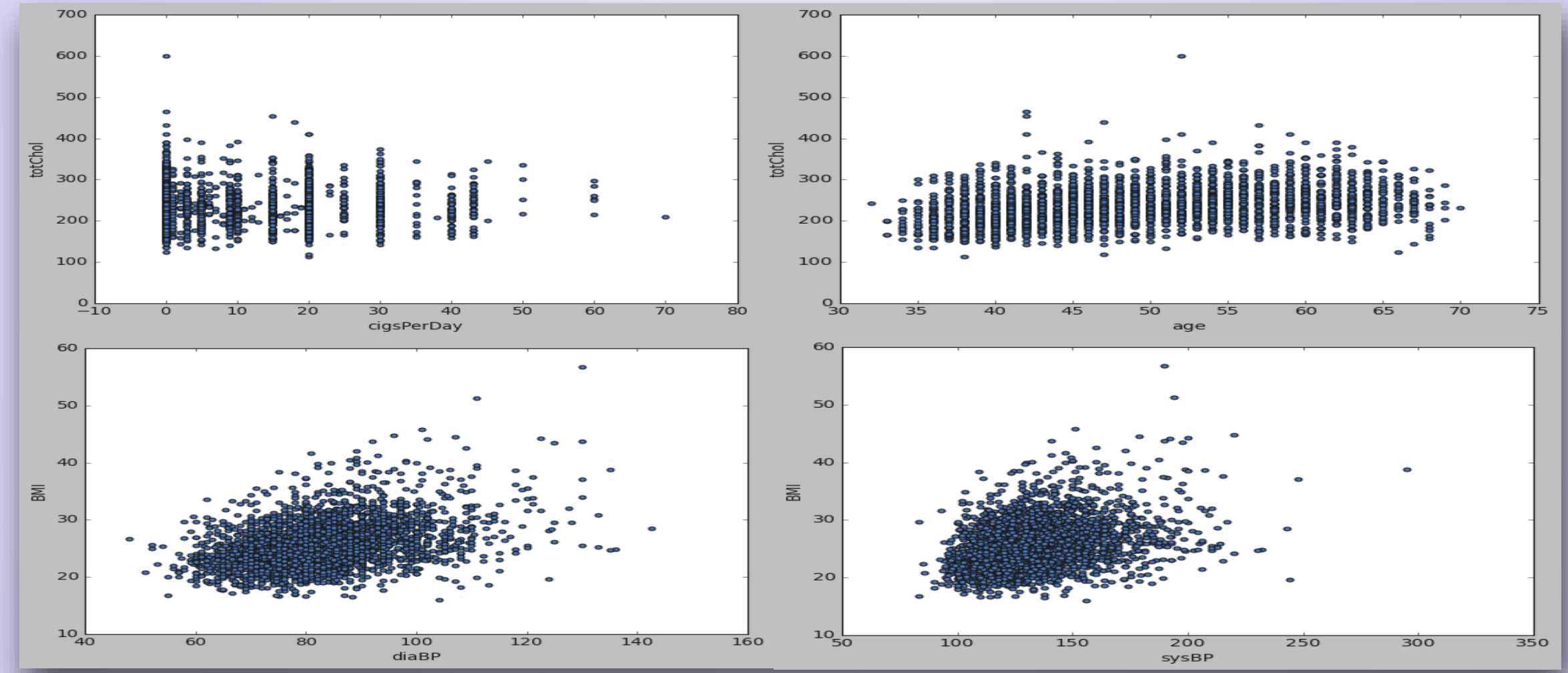
we have diastolic blood pressure (DiaBP) values, ranging from 40 to 160. Most dots are clustered between 120 and 450 total cholesterol and between 50 and 110 DiaBP, showing where most people's values fall.

Cardiovascular Health Assessment and Risk Prediction

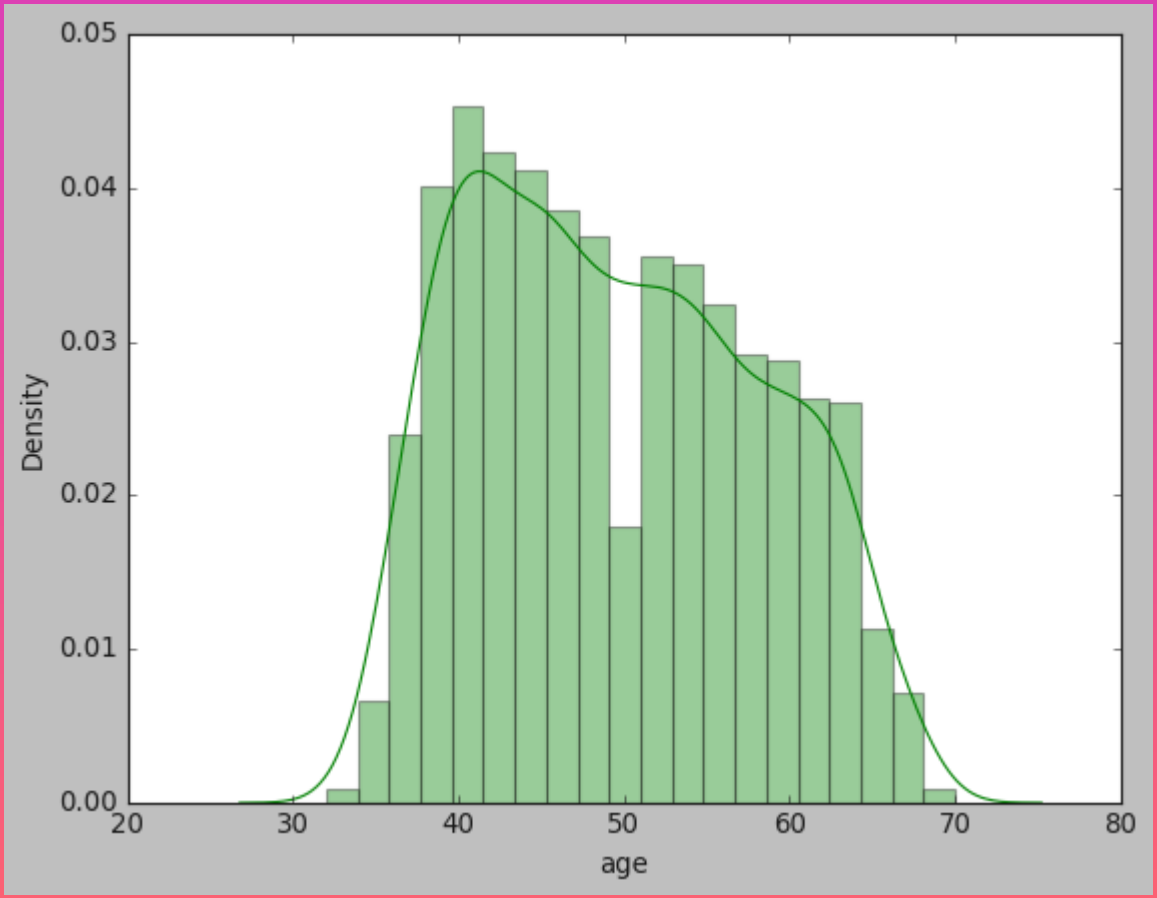


Most of the data points seem to cluster between 60 and 130 for diastolic blood pressure and between 75 and 200 for systolic blood pressure. This concentration suggests where the majority of individuals' blood pressure levels fall within the dataset.

Cardiovascular Health Assessment and Risk Prediction



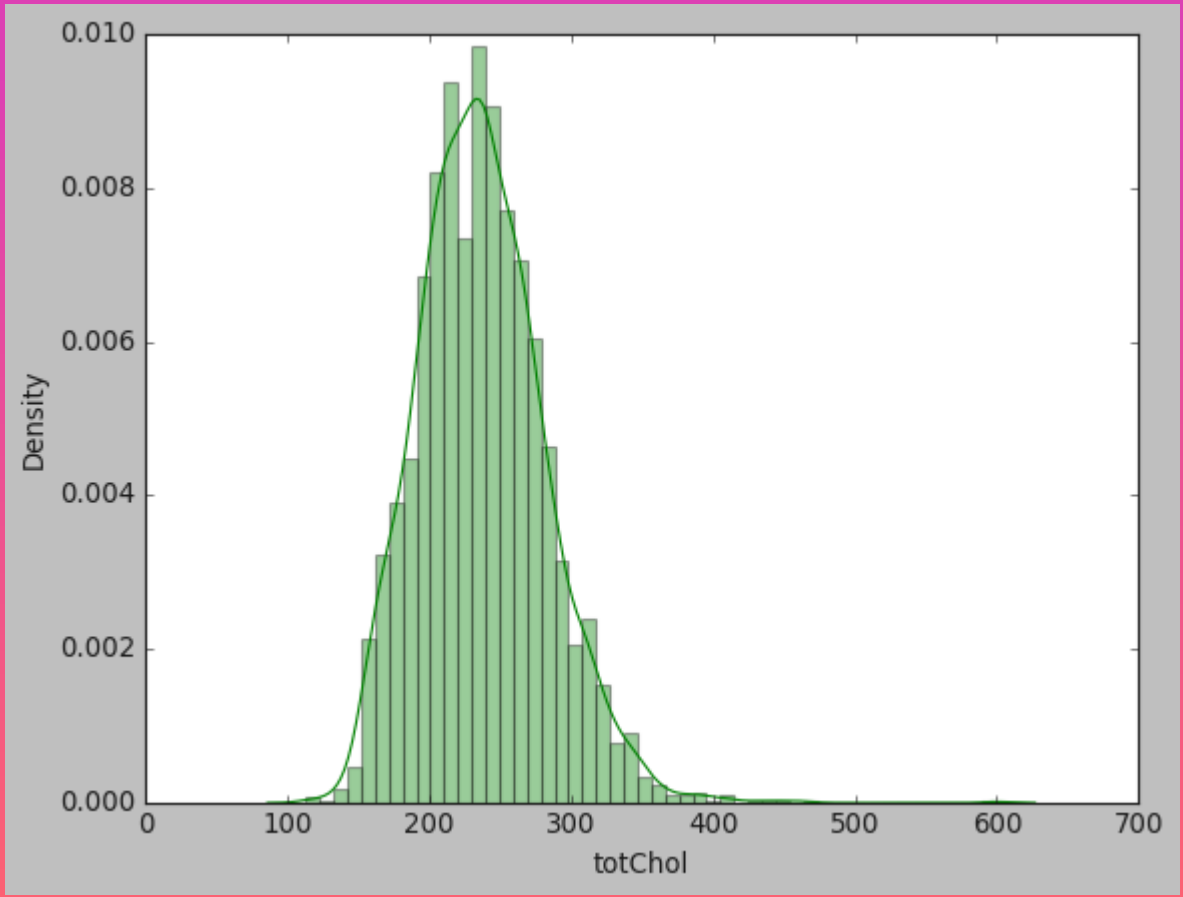
Cardiovascular Health Assessment and Risk Prediction



The histogram would likely show a concentration of data points between 30 and 70, indicating that most individuals' ages fall within this range. You can observe the shape of the distribution to determine if it's symmetric or skewed, locate the peak of the distribution (central tendency), and assess the variability of age values.

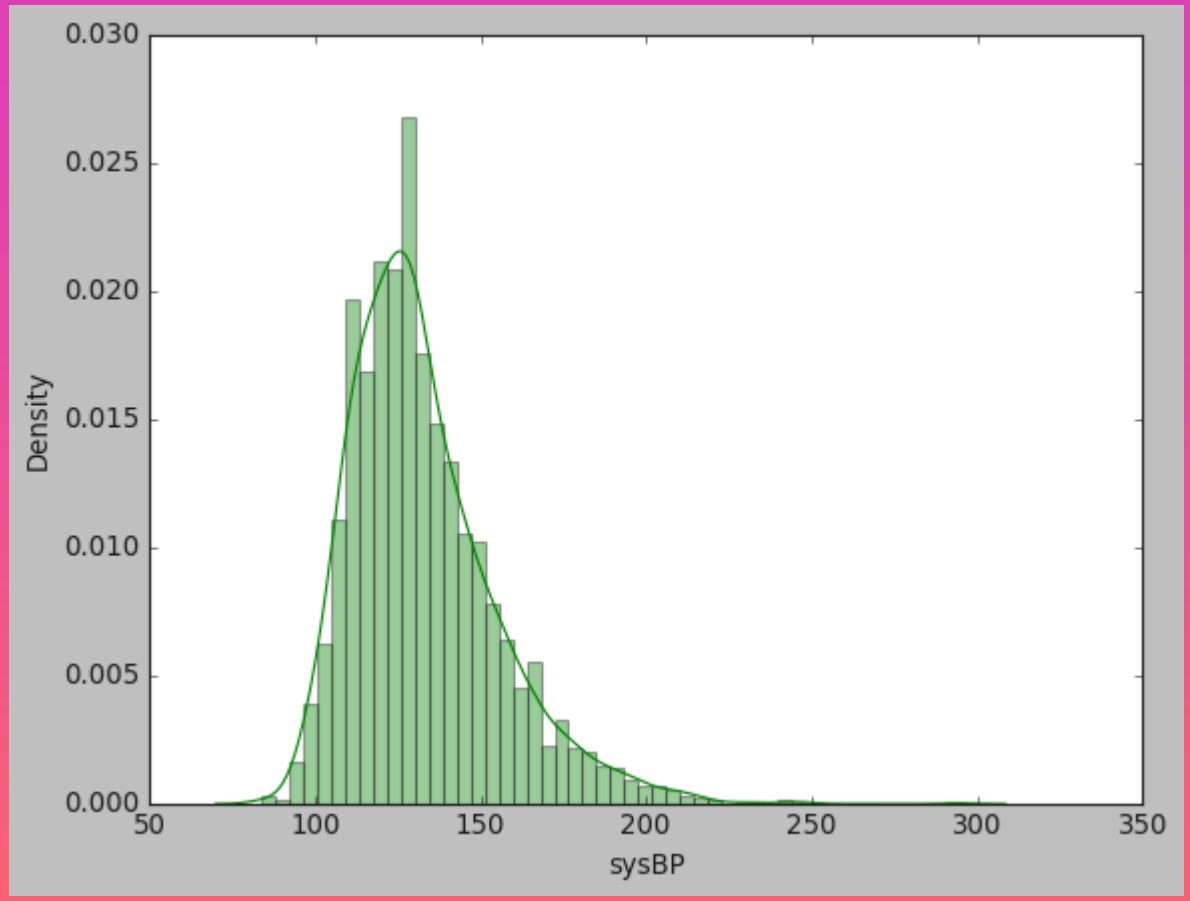
Overall, the distplot would provide a visual representation of the distribution of age values in the dataset, highlighting the prevalence of ages between 30 and 70 and offering insights into the overall age distribution.

Cardiovascular Health Assessment and Risk Prediction



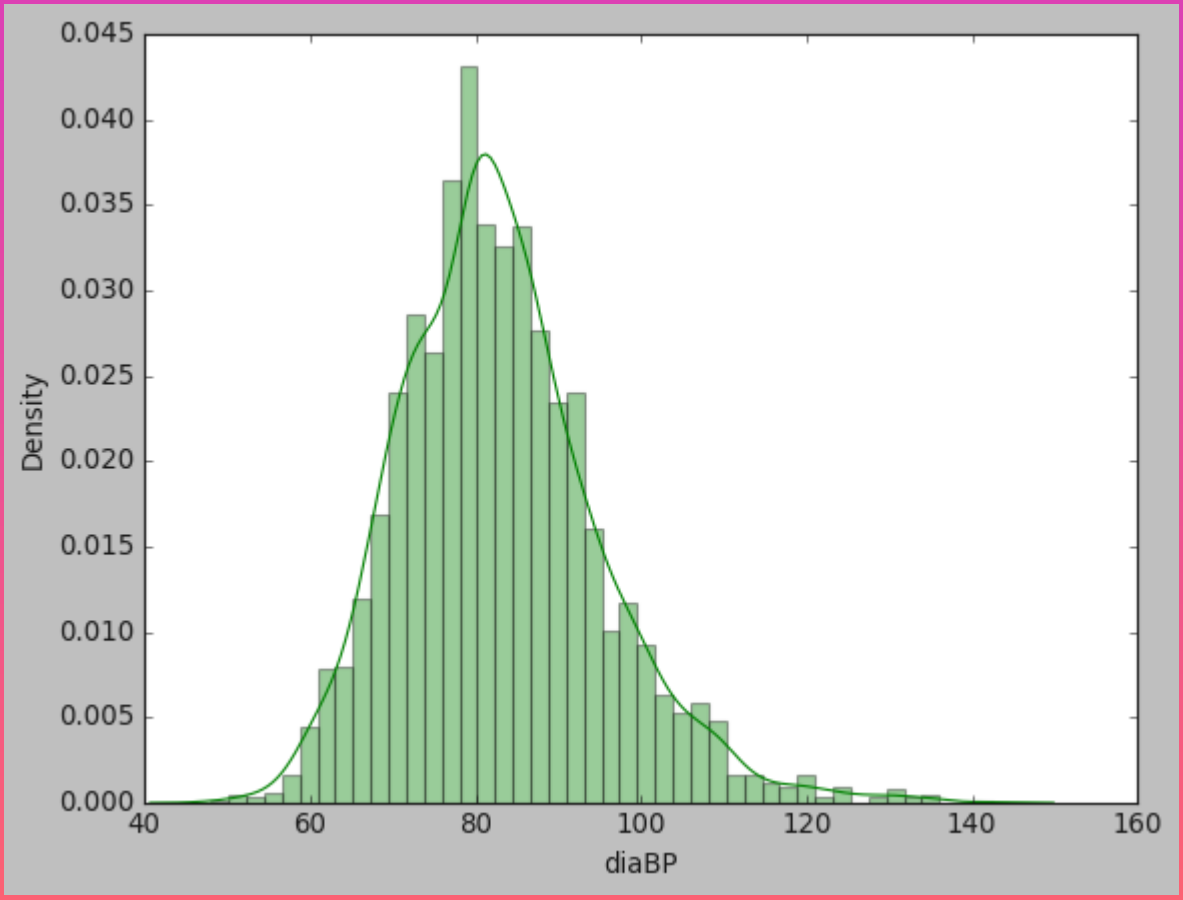
The histogram would likely show a concentration of data points between 120 and 400, indicating that most individuals' total cholesterol levels fall within this range. You can observe the shape of the distribution to determine if it's symmetric or skewed, locate the peak of the distribution (central tendency), and assess the variability of totChol values.
Overall, the distplot would provide a visual representation of the distribution of total cholesterol values in the dataset, highlighting the prevalence of values between 120 and 400 and offering insights into the overall totChol distribution.

Cardiovascular Health Assessment and Risk Prediction



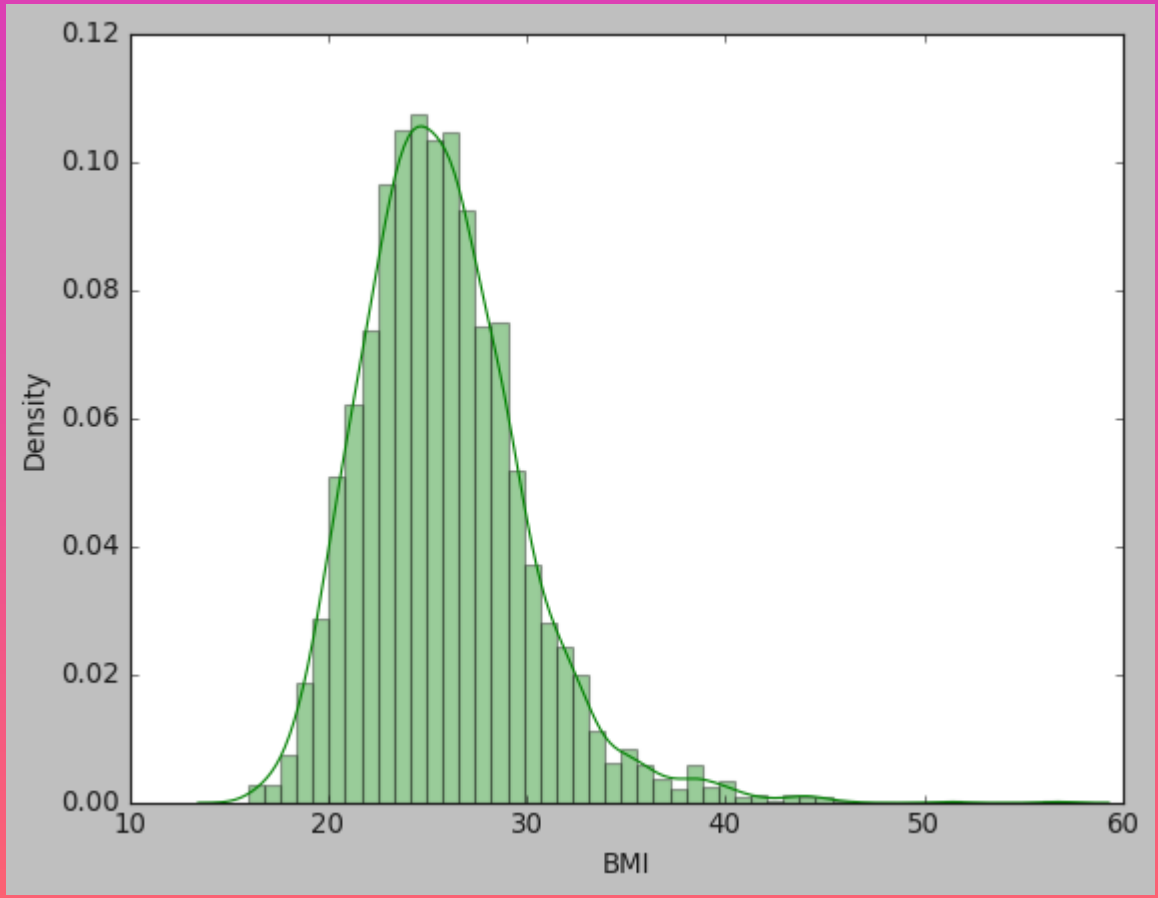
A distplot for systolic blood pressure (sysBP) between 50 and 350, with a majority falling between 100 and 200, may exhibit a right-skewed distribution. This means that the histogram would taper off to the right, with a longer tail extending towards higher systolic blood pressure values. The concentration of values towards the lower end (100-200) indicates a right skew, suggesting that while most individuals' blood pressure levels lie within this range, there are also some individuals with higher blood pressure values, contributing to the longer tail on the right side of the histogram.

Cardiovascular Health Assessment and Risk Prediction



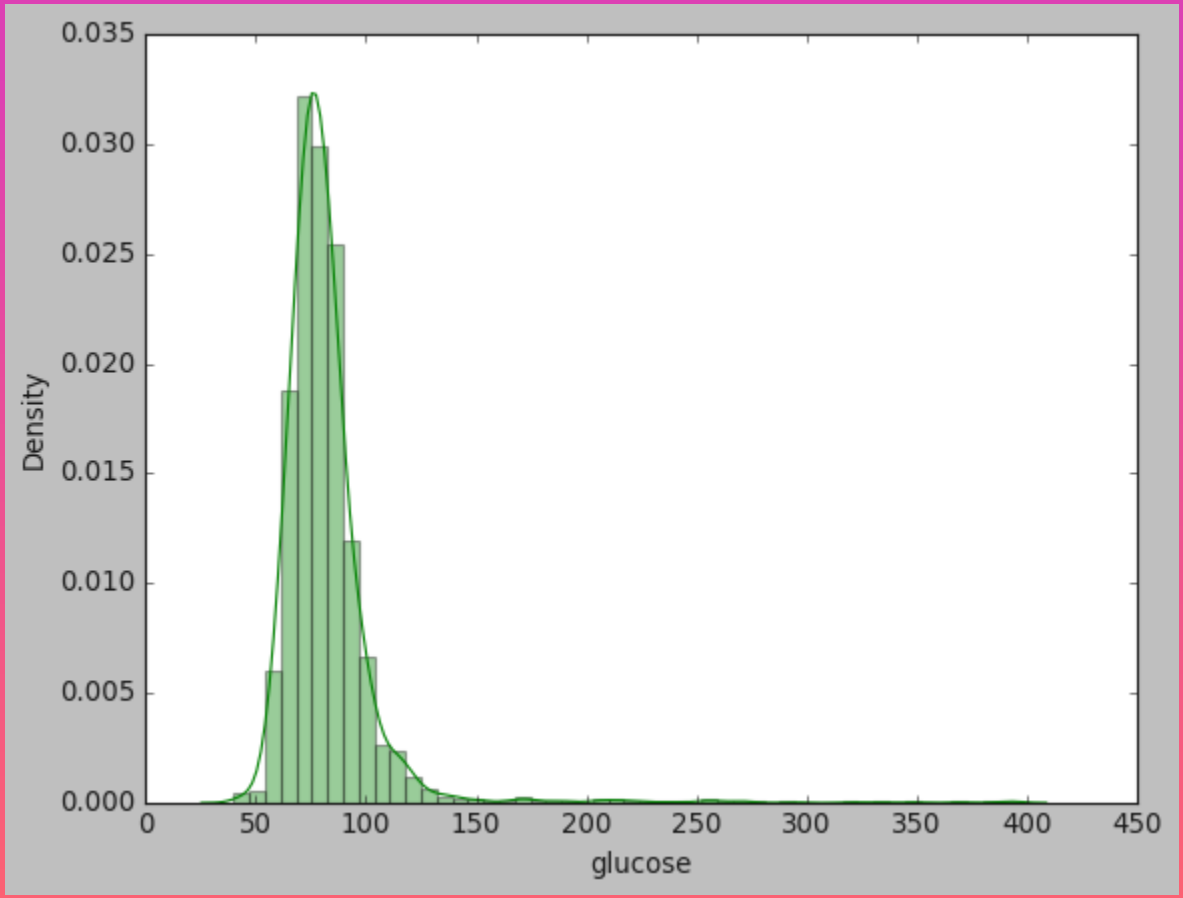
A distplot for diastolic blood pressure (DiaBP) between 40 and 160, with most values between 60 and 120, displays a histogram showing the frequency of DiaBP values. Concentration between 60 and 120 suggests most individuals' blood pressure levels fall within this range, providing insights into the overall distribution.

Cardiovascular Health Assessment and Risk Prediction



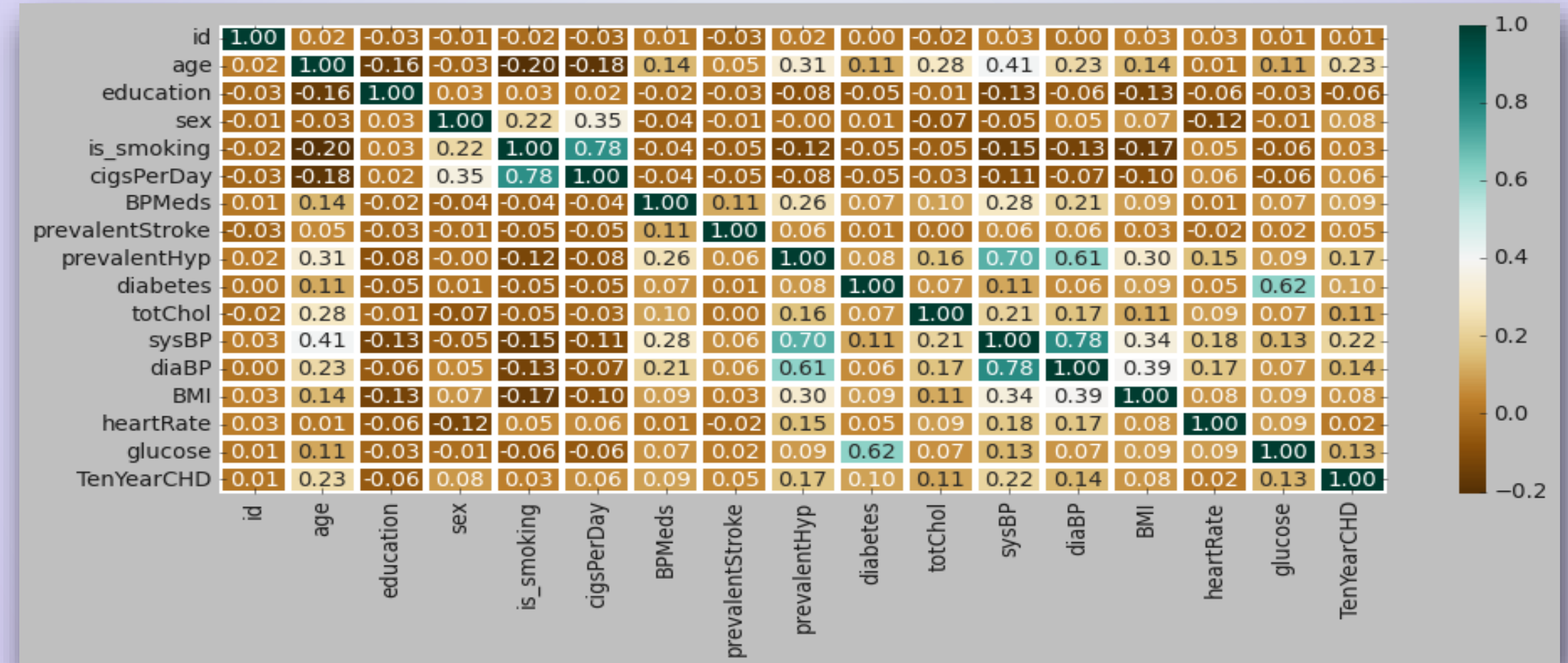
A distplot for BMI between 10 and 60, with a majority falling between 20 and 40, may exhibit a symmetric distribution or a slight right skew. If the histogram shows a longer tail towards higher BMI values, it indicates a right skew, while a symmetric distribution would show equal tapering on both sides of the peak (20-40).

Cardiovascular Health Assessment and Risk Prediction



A distplot for glucose levels between 0 and 450, with a majority falling between 50 and 150, typically shows a right-skewed distribution. This means that the histogram would taper off to the right, with a longer tail extending towards higher glucose values. The concentration between 50 and 150 indicates a right skew, suggesting that while most individuals' glucose levels lie within this range, there are also some individuals with higher glucose values, contributing to the longer tail on the right side of the histogram.

Cardiovascular Health Assessment and Risk Prediction



Cardiovascular Health Assessment and Risk Prediction

Introduction

EDA Of Data

Feature Engineering

Feature engineering involves encoding categorical variables, generating interaction terms, normalizing numerical features, and handling missing values. Domain-specific insights guide the creation of additional features. Feature selection techniques identify the most predictive variables. These steps optimize the dataset for CHD risk prediction modeling.

Discussion and Conclusion

Cardiovascular Health Assessment and Risk Prediction

Feature Engineering

Label Encoding

Convert categorical variables into numerical format using label encoding, enabling the model to interpret categorical data. For example, encoding binary variables like 'sex' (male/female) as 0 and 1.

Random Over Sampler

Min Max Scaler

Principal Component Analysis

Cardiovascular Health Assessment and Risk Prediction

Feature Engineering

Label Encoding

Random Over Sampler

Address class imbalance by randomly duplicating minority class instances or generating synthetic data points. This helps prevent the model from being biased towards the majority class, ensuring balanced training data.

Min Max Scaler

Principal Component Analysis

Cardiovascular Health Assessment and Risk Prediction

Feature Engineering

Label Encoding

Random Over Sampler

Min Max Scaler

Normalize numerical features to a predefined range (e.g., [0, 1]) to ensure consistency and comparability across different features. This scaling technique preserves the relationship between features while preventing features with larger scales from dominating the model.

Principal Component Analysis

Cardiovascular Health Assessment and Risk Prediction

Feature Engineering

Label Encoding

Random Over Sampler

Min Max Scaler

Principal Component Analysis

Reduce dimensionality by transforming high-dimensional data into a lower-dimensional space while preserving most of the variance. PCA helps mitigate the curse of dimensionality, improves computational efficiency, and removes redundant or noisy features.

Cardiovascular Health Assessment and Risk Prediction

Introduction

EDA Of Data

Feature Engineering

Discussion and Conclusion

The analysis highlights key factors influencing CHD risk, informing preventive strategies. Future research could explore advanced predictive models and additional risk factors. Overall, understanding CHD risk factors is crucial for effective prevention and early intervention, ultimately improving public health outcomes.

Cardiovascular Health Assessment and Risk Prediction

1. **Random Forest Classifier** achieved the highest accuracy of 92.86%. It exhibited a balanced performance with 437 true negatives (TN), 486 true positives (TP), 59 false positives (FP), and 12 false negatives (FN), indicating good predictive power and generalization capability.
2. **Bagging Classifier** also demonstrated high accuracy (91.85%) with a slightly different confusion matrix, indicating similar performance to Random Forest Classifier. It correctly classified 432 TN, 481 TP, 64 FP, and 17 FN instances.
3. **Gradient Boosting Classifier** yielded a relatively lower accuracy of 73.04% compared to the previous classifiers. It showed a higher number of false positives (162) and false negatives (106), suggesting potential overfitting or difficulty in capturing the underlying patterns in the data.
4. **AdaBoost Classifier** also exhibited lower accuracy (69.11%) with a higher number of false positives (165) and false negatives (142). It may struggle with classifying instances accurately, possibly due to the complexity of the dataset or insufficient model training.
5. **Decision Tree Classifier** achieved a commendable accuracy of 90.34%, with a balanced confusion matrix. It showed good performance in classifying instances, albeit slightly lower than RandomForestClassifier and BaggingClassifier.
6. **Logistic Regression** with balanced class weights and the "ovr" multi-class strategy demonstrated the lowest accuracy (66.70%) among the classifiers. It struggled with false positives (170) and false negatives (161), indicating challenges in accurately predicting class labels.

Cardiovascular Health Assessment and Risk Prediction

In conclusion, Random Forest Classifier and Bagging Classifier performed the best in terms of accuracy and robustness, followed by Decision Tree Classifier. Gradient Boosting Classifier and AdaBoost Classifier showed lower performance, possibly due to overfitting or difficulty in capturing the data's complexity. Logistic Regression performed the poorest among the classifiers, suggesting limitations in handling the dataset's characteristics effectively. Further optimization and fine-tuning may enhance the performance of classifiers, ultimately improving predictive accuracy for CHD risk assessment.

Thank You

Cardiovascular Health Assessment and Risk Prediction

Yash Sharma
Data Scientist & Assistant Professor
Exp: 8+Year
Mob:9691173377
yash.versatile@gmail.com