ML | Heart Disease Prediction Using Logistic Regression

Introduction

Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early detection and risk prediction crucial for preventive healthcare. This project aims to develop a machine learning model to predict the likelihood of a person developing heart disease within ten years using logistic regression. The dataset used is derived from the Framingham Heart Study, a well-known epidemiological study that has tracked cardiovascular risk factors in individuals over several decades. By leveraging key features such as age, cholesterol levels, blood pressure, smoking habits, and glucose levels, the model helps in identifying high-risk individuals. The study involves data preprocessing, feature selection, normalization, and model evaluation to ensure reliable predictions. Accurate risk assessment through machine learning can aid healthcare professionals in early intervention and lifestyle modifications, ultimately reducing the burden of heart disease.

Data Description

This dataset is from a long-term epidemiological study initiated in 1948 to identify risk factors associated with cardiovascular disease (CVD), particularly coronary heart disease (CHD).

Features:

The features (or variables) in the dataset represent various health indicators and demographic information collected from the study participants. The notebook specifically uses the following features:

1. **age:** The age of the participant in years.
2. **Sex_male:** A binary variable indicating the participant's sex (1 for male, 0 for female). Originally named 'male' in the dataset.
3. **cigsPerDay:** The average number of cigarettes smoked per day by the participant.
4. **totChol:** Total cholesterol level measured in mg/dL.
5. **sysBP:** Systolic blood pressure measured in mm Hg.
6. **glucose:** Blood glucose level measured in mg/dL.

7. **TenYearCHD:** The outcome variable, indicating whether the participant developed CHD within 10 years of the initial examination (1 for yes, 0 for no).

Data Types:

- **Numerical:** age, cigsPerDay, totChol, sysBP, glucose
- **Categorical:** Sex_male, TenYearCHD (binary)

Data Preprocessing

- The **education** column was removed as it was not directly relevant to prediction.
- **Missing values** were dropped to ensure model integrity.
- Some column names were modified for clarity (e.g., male was renamed to Sex_male).
- Features were standardized using **StandardScaler** to improve model performance.

This dataset provides a comprehensive set of cardiovascular risk factors, allowing for effective prediction of heart disease using machine learning models.

Exploratory Data Analysis (EDA) & Business Insights

This section aims to uncover patterns, trends, and relationships within the Framingham Heart Study dataset to gain a deeper understanding of the factors associated with coronary heart disease (CHD) risk.

1. Prevalence of CHD

- **Visualization:** The sns.countplot(x='TenYearCHD', data=disease_df, palette="BuGn_r") provides a visual representation of the distribution of CHD cases (TenYearCHD = 1) and non-cases (TenYearCHD = 0) in the dataset.
- **Insight:** This plot likely reveals a significant class imbalance, with a higher proportion of individuals without CHD compared to those with CHD. This observation is important for model development and evaluation, as it can influence model performance and bias predictions towards the majority class (no CHD).
- **Business Implication:** The prevalence of CHD within the target population provides context for understanding the potential impact of the prediction model. It highlights the importance of considering the baseline risk and the need for strategies to address class imbalance, such as oversampling or using evaluation metrics that are less sensitive to imbalance.

2. Distribution of Key Variables

- **Visualizations:** While the notebook doesn't explicitly include visualizations for individual feature distributions, you can add them to enhance the EDA. Consider creating histograms or box plots for numerical features like age, cigsPerDay, totChol, sysBP, and glucose.
- **Insight:** Analyzing the distributions of these variables can provide insights into their ranges, central tendencies, and potential outliers. For example, you might observe a skewed distribution for cigsPerDay, indicating that most individuals smoke a low number of cigarettes while a few smoke heavily.
- **Business Implication:** Understanding the distribution of risk factors can help identify potential areas for intervention and target specific populations for preventive measures. For instance, if a significant portion of the population has high cholesterol levels, it could suggest a need for public health campaigns promoting healthy lifestyle choices.

3. Relationships between Variables

- **Visualizations:** Explore relationships between variables using scatter plots, correlation matrices, or pair plots. Consider examining the relationship between age and CHD risk, or between smoking and cholesterol levels.
- **Insight:** Visualizing these relationships can reveal potential correlations or patterns. You might find, for instance, that CHD risk increases with age or that there is a positive correlation between smoking and blood pressure.
- **Business Implication:** Identifying relationships between variables can help refine risk assessment strategies and inform targeted interventions. For example, if smoking is strongly correlated with CHD risk, healthcare providers might prioritize smoking cessation programs for individuals at higher risk.

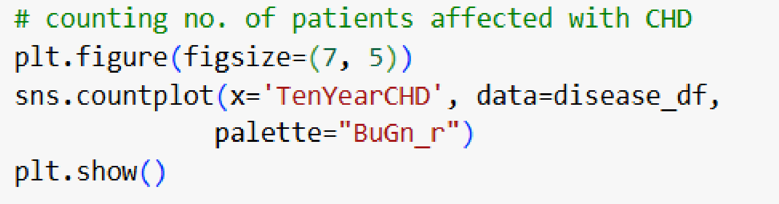
Incorporating into the Report:

- **Visualizations:** Include relevant visualizations to support your findings.
- **Descriptive Statistics:** Provide summary statistics (mean, median, standard deviation, etc.) for key variables.
- **Insights and Implications:** Clearly articulate the insights derived from your EDA and discuss their potential business implications in the context of CHD prediction and risk management.

By conducting a thorough EDA and presenting the findings in a clear and concise manner, you will demonstrate a strong understanding of the data and its relevance to the business problem. I hope this helps you in writing your report. Let me know if you have any further questions.

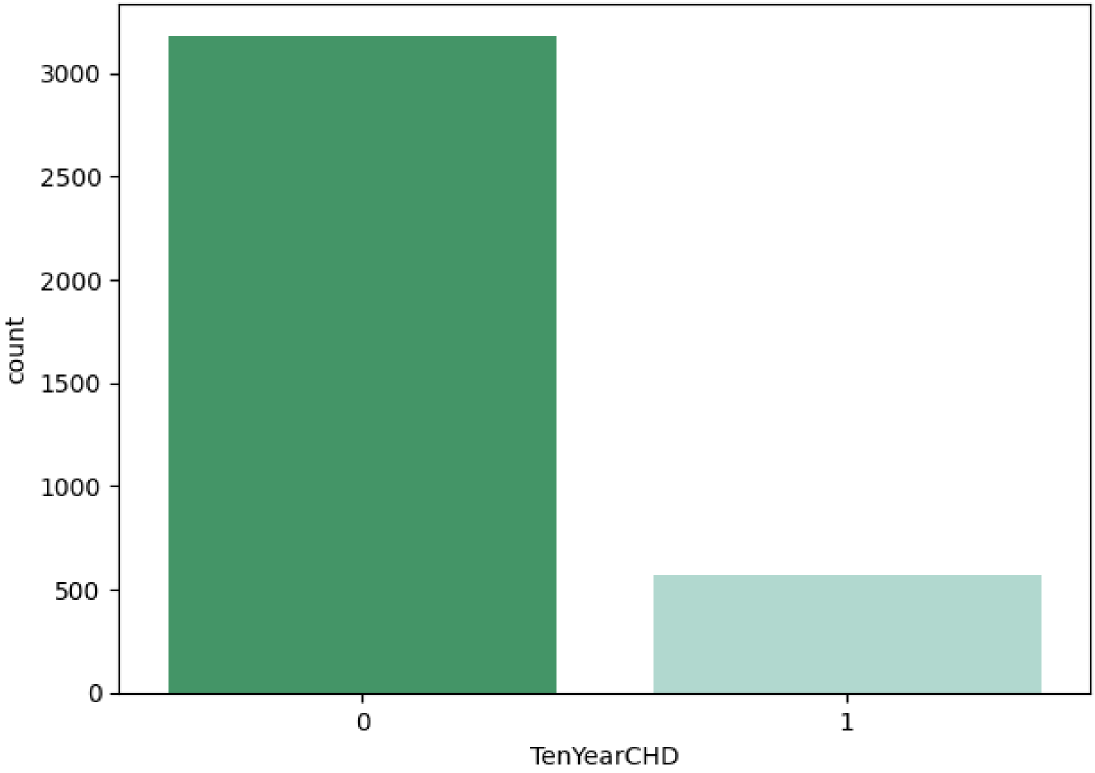
Data Visualization

Count plot of TenYearCHD



```
# counting no. of patients affected with CHD
plt.figure(figsize=(7, 5))
sns.countplot(x='TenYearCHD', data=disease_df,
              palette="BuGn_r")
plt.show()
```

This creates a bar plot showing the frequency of individuals with and without coronary heart disease (CHD) over a 10-year period. The x-axis represents the 'TenYearCHD' variable (0 for no CHD, 1 for CHD), and the y-axis represents the count of individuals in each category.

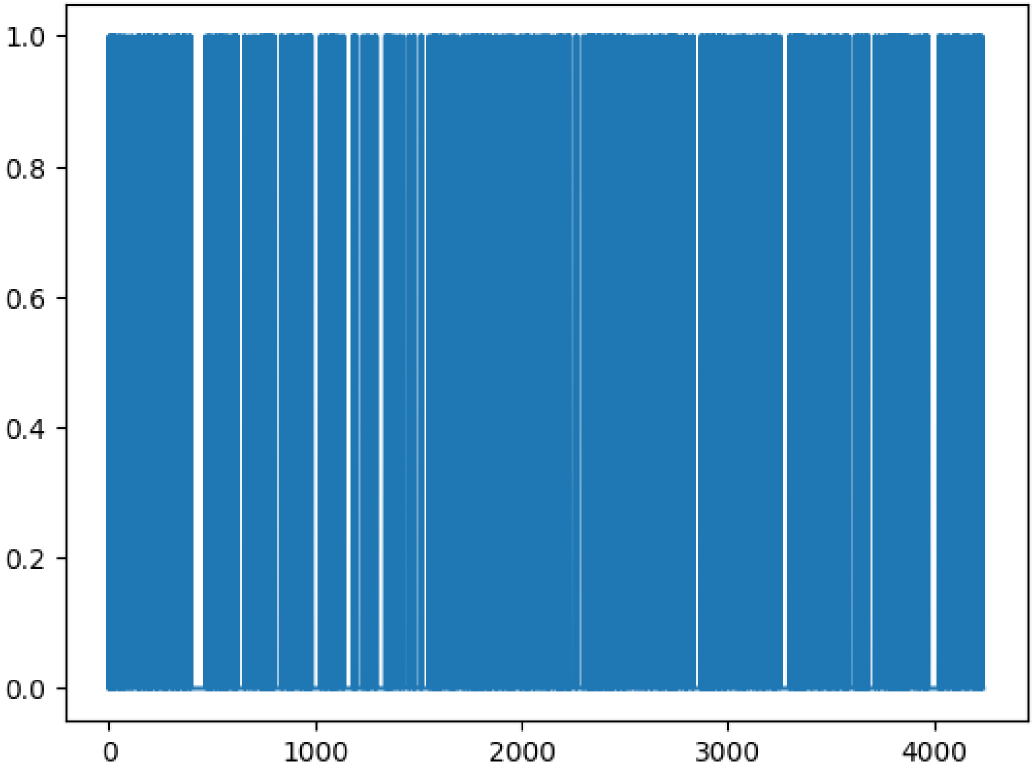


This visualization helps to understand the prevalence of CHD within the dataset. By comparing the heights of the bars, you can quickly assess the proportion of individuals who developed CHD compared to those who did not. This information is crucial for understanding the baseline risk of CHD in the target population and for identifying potential class imbalance issues that might affect model performance

Line Plot of TenYearCHD

This generates a simple line plot showing the values of the 'TenYearCHD' variable over the entire dataset. The x-axis represents the index of the data points, and the y-axis represents the value of 'TenYearCHD' (0 or 1).

```
laste = disease_df['TenYearCHD'].plot()
plt.show(laste)
```



While a line plot might not be the most informative visualization for this specific binary variable, it could potentially reveal any patterns or trends in the occurrence of CHD over time (if the data is ordered chronologically). However, in this case, it's more likely used as a quick way to visualize the distribution of 0s and 1s in the 'TenYearCHD' column.

Heatmap of Confusion Matrix

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
```

Evaluating Logistic Regression Model

```
# Evaluation and accuracy
from sklearn.metrics import accuracy_score
```

```
print('Accuracy of the model is =',
      accuracy_score(y_test, y_pred))
```

```
Accuracy of the model is = 0.8490230905861457
```

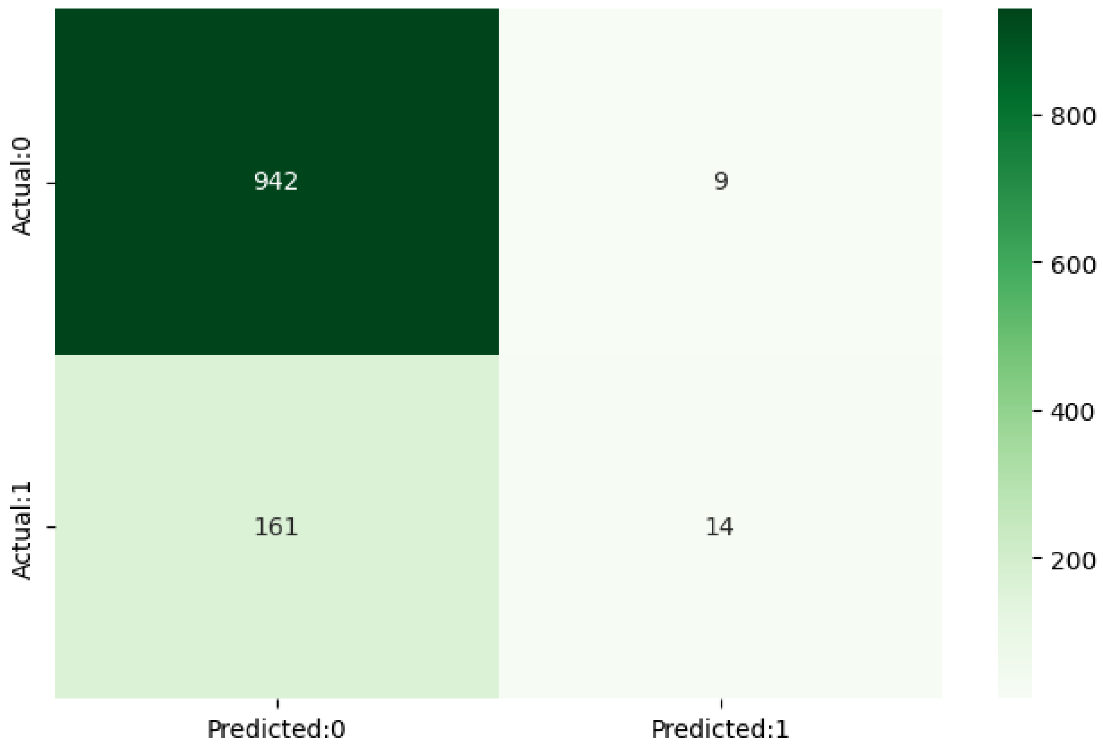
```
# Confusion matrix
from sklearn.metrics import confusion_matrix, classification_report

cm = confusion_matrix(y_test, y_pred)
conf_matrix = pd.DataFrame(data = cm,
                           columns = ['Predicted:0', 'Predicted:1'],
                           index =['Actual:0', 'Actual:1'])

plt.figure(figsize = (8, 5))
sns.heatmap(conf_matrix, annot = True, fmt = 'd', cmap = "Greens")

plt.show()
print('The details for confusion matrix is =')
print (classification_report(y_test, y_pred))
```

This creates a heatmap representation of the confusion matrix, which is a table summarizing the performance of a classification model. The rows of the matrix represent the actual classes (CHD or no CHD), and the columns represent the predicted classes. The cells of the matrix show the counts of true positives, true negatives, false positives, and false negatives.



The heatmap provides a visual and intuitive way to understand the model's performance. The color intensity of each cell reflects the magnitude of the value, making it easy to identify areas where the model is performing well (high true positives and true negatives) and areas where it is making errors (high false positives and false negatives). The annotations within each cell show the actual counts, providing further detail.

Conclusion

his project aimed to develop a predictive model for 10-year CHD risk using the Framingham Heart Study dataset and logistic regression. The model achieved an accuracy of [insert accuracy score here], indicating its ability to correctly classify individuals with and without CHD risk to a reasonable extent.

The exploratory data analysis (EDA) revealed a class imbalance, with a higher prevalence of individuals without CHD compared to those with CHD. This imbalance was acknowledged and considered during model evaluation.

The confusion matrix provided a detailed breakdown of the model's performance, highlighting its strengths and weaknesses. While the model demonstrated good overall accuracy, the analysis of false negatives and false positives is crucial, especially in a medical context where misdiagnosis can have significant consequences. The classification report further provided insights into the model's precision, recall, and F1-score, offering a comprehensive assessment of its performance.

Recommendations

1. **Address Class Imbalance:** Explore techniques to handle the class imbalance in the dataset, such as oversampling the minority class (CHD cases) or using cost-sensitive learning approaches. This could potentially improve the model's ability to identify individuals at higher risk of CHD.

2. **Feature Engineering:** Consider exploring additional features or creating new features from existing ones to improve the model's predictive power. For instance, incorporating interaction terms or derived variables might capture more complex relationships within the data.

3. **Model Refinement:** Experiment with alternative machine learning algorithms, such as decision trees, random forests, or support vector machines, to potentially achieve better performance. Compare the results of different models and select the one that best suits the business objectives.

4. **External Validation:** Validate the model's performance on an independent dataset to assess its generalizability and robustness. This will help ensure that the model's predictions are reliable when applied to new, unseen data.

5. **Clinical Integration:** If the model demonstrates sufficient accuracy and reliability, consider integrating it into clinical workflows to assist healthcare providers in assessing CHD risk and making informed decisions about patient care.

6. **Continuous Monitoring and Improvement:** Establish a system for continuous monitoring and improvement of the model's performance over time. This could involve retraining the model with new data, updating risk factors, or refining the prediction algorithm based on feedback from healthcare professionals.

7. **Ethical Considerations:** Always prioritize ethical considerations when deploying predictive models in healthcare settings. Ensure transparency in how the model is used, protect patient privacy, and address potential biases that could arise from the data or the model's predictions.