Experiment 10 Case Study

Title: Predicting Heart Disease Risk Using Machine Learning

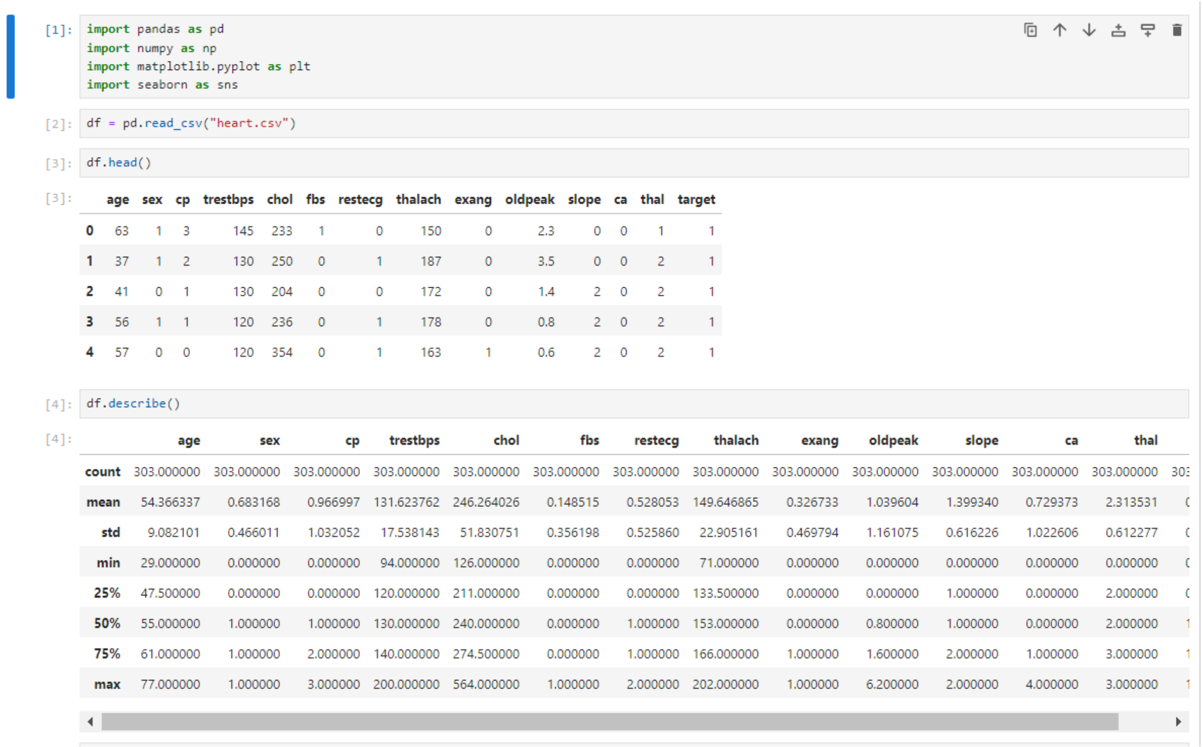
- **Abstract:** Heart disease is a significant public health concern globally, emphasizing the importance of accurate risk assessment for early intervention and prevention. This case study explores the application of machine learning algorithms in predicting heart disease risk based on patient demographics, medical attributes, and lifestyle factors. Through comprehensive data preprocessing, feature selection, model training, and evaluation, this study aims to develop a robust predictive model to assist healthcare professionals in identifying individuals at heightened risk of heart disease, thereby enabling targeted interventions and improving health outcomes.

- **Introduction:** Heart disease, including conditions such as coronary artery disease, heart failure, and arrhythmias, remains a leading cause of morbidity and mortality worldwide. Despite advancements in medical science and technology, the accurate prediction of heart disease risk remains a challenge. Traditional risk assessment methods rely on clinical factors and medical history, which may lack granularity and predictive power. In contrast, machine learning offers the potential to integrate diverse data sources and develop personalized risk prediction models. In this case study, we leverage machine learning techniques to develop a predictive model for heart disease risk assessment, utilizing a comprehensive dataset encompassing various patient attributes and cardiac health outcomes.

- **Case Description:** The dataset utilized in this case study comprises anonymized patient data collected from multiple healthcare facilities. It encompasses a wide range of features, including demographic information (age, gender), physiological parameters (blood pressure, cholesterol levels), medical history (hypertension, diabetes), lifestyle factors (smoking habits), and family history of heart disease. The target variable indicates the presence or absence of heart disease. Prior to analysis, the dataset undergoes meticulous preprocessing to handle missing values, normalize numerical features, and encode categorical variables, ensuring data quality and consistency.

- **Methodology:**

1. **Data Preprocessing:**
   - Missing value imputation: Employ techniques such as mean imputation, median imputation, or predictive modeling to handle missing data.
   - Feature scaling: Normalize numerical features to a standard scale (e.g., z-score normalization) to facilitate model convergence and interpretation.
   - Encoding categorical variables: Transform categorical attributes into numerical representations using methods like one-hot encoding or label encoding.



```
[1]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns

[2]: df = pd.read_csv("heart.csv")

[3]: df.head()
```

[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

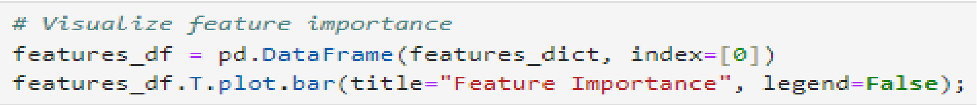
```
[4]: df.describe()
```

[4]:

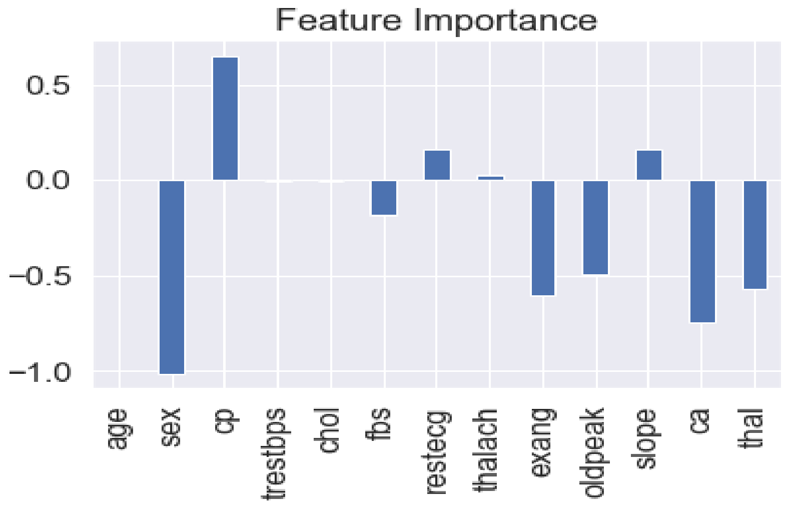
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	(
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	(
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	(
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	(
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1

2. **Feature Selection:**

- Conduct feature selection to identify informative attributes relevant to heart disease prediction.
- Utilize techniques such as correlation analysis, feature importance scores (e.g., based on tree-based models), or recursive feature elimination to identify and retain the most relevant features.

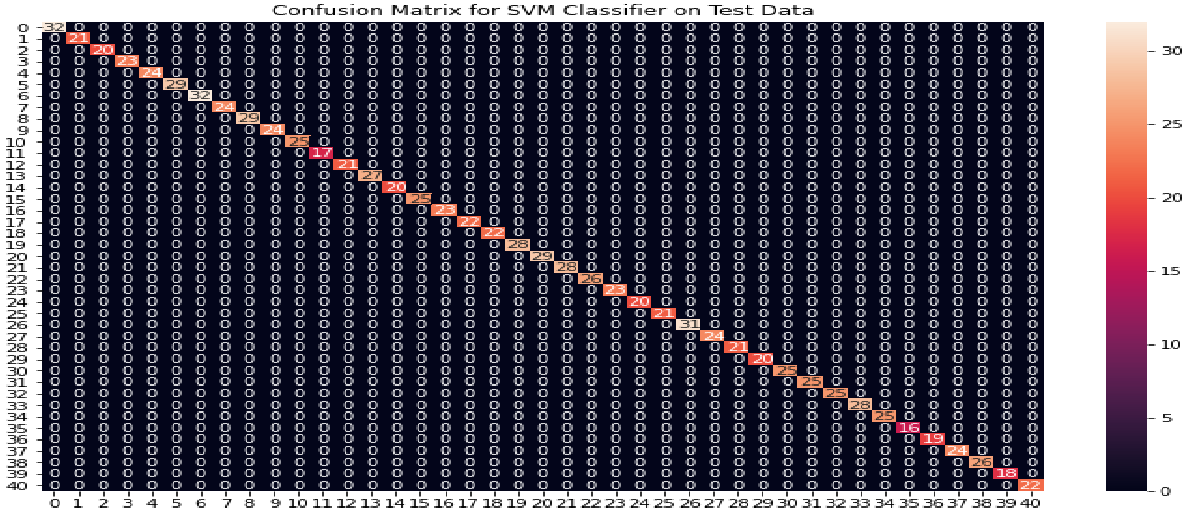


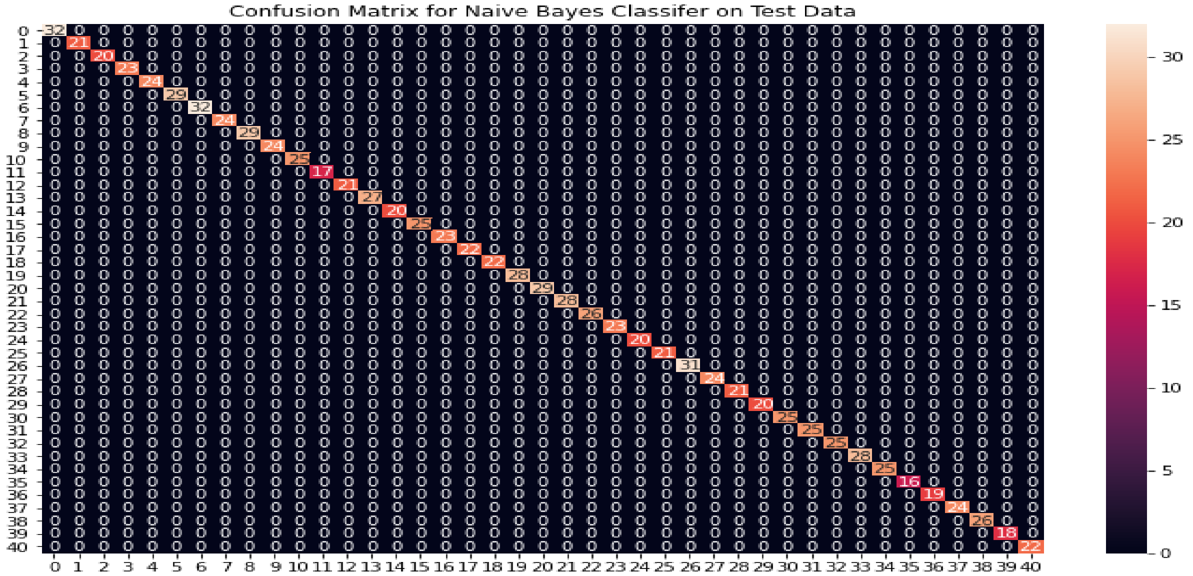
```
# Visualize feature importance
features_df = pd.DataFrame(features_dict, index=[0])
features_df.T.plot.bar(title="Feature Importance", legend=False);
```

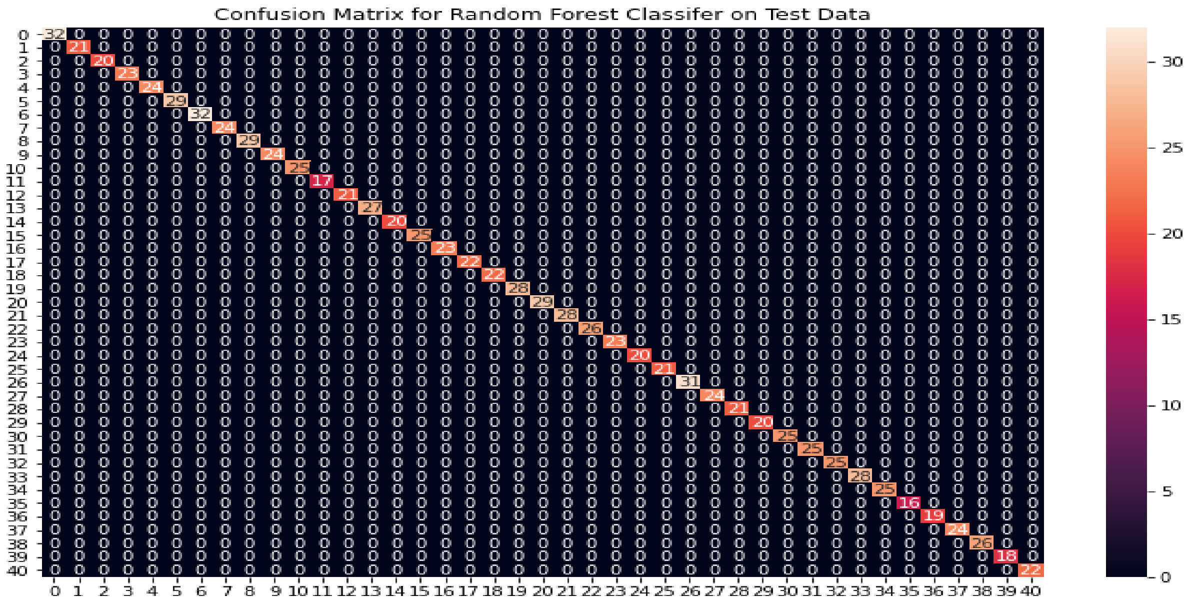


3. Model Training:

- Split the dataset into training, validation, and testing sets to ensure robust model evaluation.
- Train various machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, gradient boosting, and neural networks.
- Optimize hyperparameters using techniques such as grid search, randomized search, or Bayesian optimization to enhance model performance.

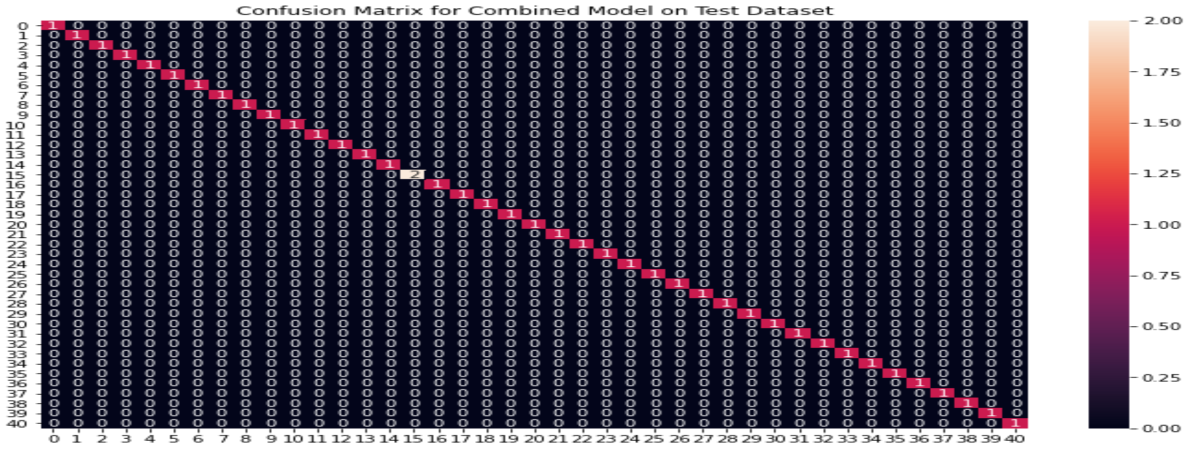






4. **Model Evaluation:**

- Evaluate the performance of each model using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (ROC-AUC), and confusion matrix analysis.
- Conduct cross-validation to assess model generalizability and mitigate overfitting.
- Perform statistical significance testing to compare the performance of different algorithms and identify the most effective approach for heart disease prediction.



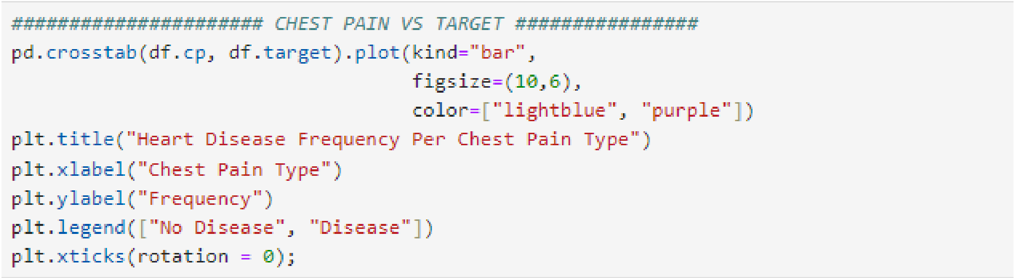
❖ **Analysis:** The analysis reveals insightful findings regarding the predictive power of different features and machine learning algorithms in identifying individuals at risk of heart disease. Notably, attributes such as age, cholesterol levels, blood pressure, and smoking status emerge as significant predictors of heart disease risk. Machine learning models, particularly ensemble methods like random forests and gradient boosting, demonstrate superior performance compared to traditional statistical approaches, achieving high accuracy and robustness in predicting heart disease outcomes. Feature importance analysis elucidates the relative contributions of individual features to the predictive model's performance, providing valuable insights into the underlying mechanisms of heart disease development and progression.

❖ **Solution/Recommendations:** Based on the analysis results, we recommend the deployment of an ensemble learning approach, such as a random forest or gradient boosting classifier, as the preferred predictive model for heart disease risk assessment in clinical practice. These models offer a balance between predictive accuracy, interpretability, and computational efficiency, making them well-suited for real-world applications. Healthcare providers can leverage the predictive model output to stratify patients based on their estimated risk of heart disease and tailor preventive interventions accordingly. Additionally, continuous monitoring and refinement of the predictive model using updated data and feedback from clinical practice are essential to ensure its ongoing relevance and effectiveness in improving cardiovascular health outcomes.

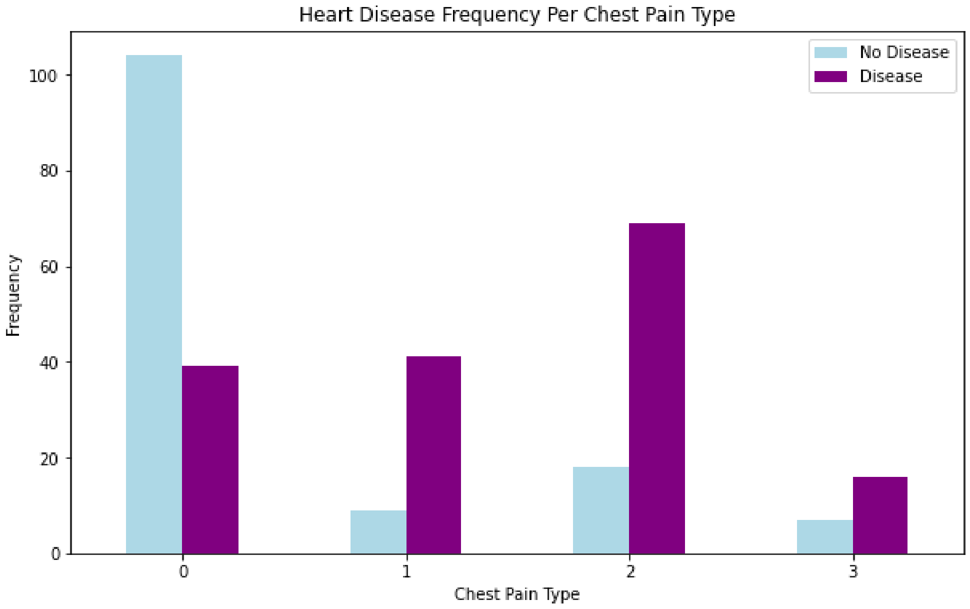
❖ **Implementation:** The implementation of the predictive model entails several critical steps, including integration into existing electronic health record systems, development of user-friendly interfaces for healthcare providers, and establishment of protocols for model validation and updating. Training sessions and educational materials should be provided to healthcare staff to familiarize them with the model's capabilities and limitations. Furthermore, collaboration with regulatory bodies and ethical review boards is necessary to ensure compliance with data privacy regulations and ethical guidelines governing the use of patient data for research and clinical purposes.

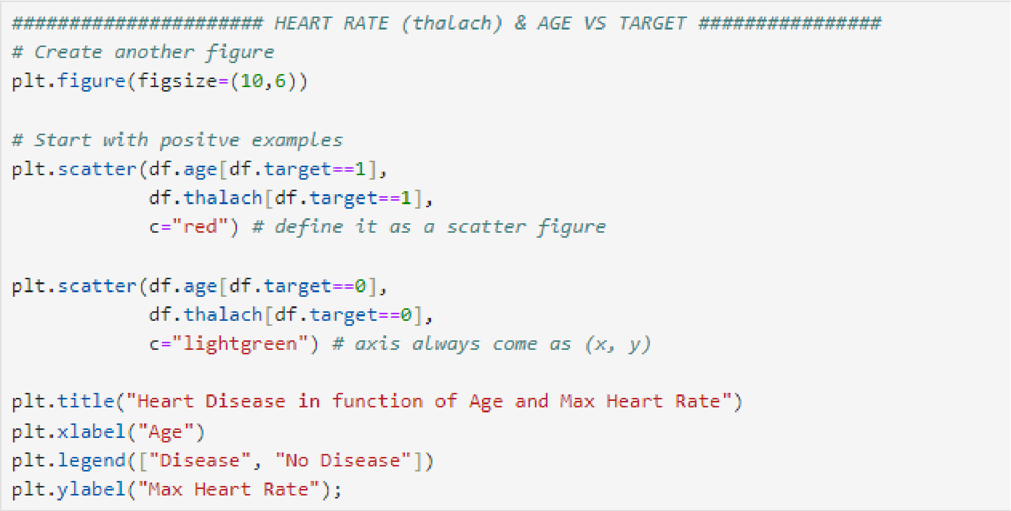
❖ **Outcomes/Evaluation:** The successful implementation of the predictive model in clinical practice facilitates more proactive and personalized approaches to heart disease prevention and management. By accurately identifying individuals at elevated risk of heart disease, healthcare providers can intervene with targeted interventions, lifestyle modifications, and pharmacological therapies, thereby mitigating the incidence and severity of cardiovascular events. Continuous evaluation and refinement of the model based on real-world data and clinical feedback are paramount to its long-term sustainability and effectiveness in improving patient outcomes.

❖ Outputs:



```
###################### CHEST PAIN VS TARGET ################
pd.crosstab(df.cp, df.target).plot(kind="bar",
                                   figsize=(10,6),
                                   color=["lightblue", "purple"])
plt.title("Heart Disease Frequency Per Chest Pain Type")
plt.xlabel("Chest Pain Type")
plt.ylabel("Frequency")
plt.legend(["No Disease", "Disease"])
plt.xticks(rotation = 0);
```



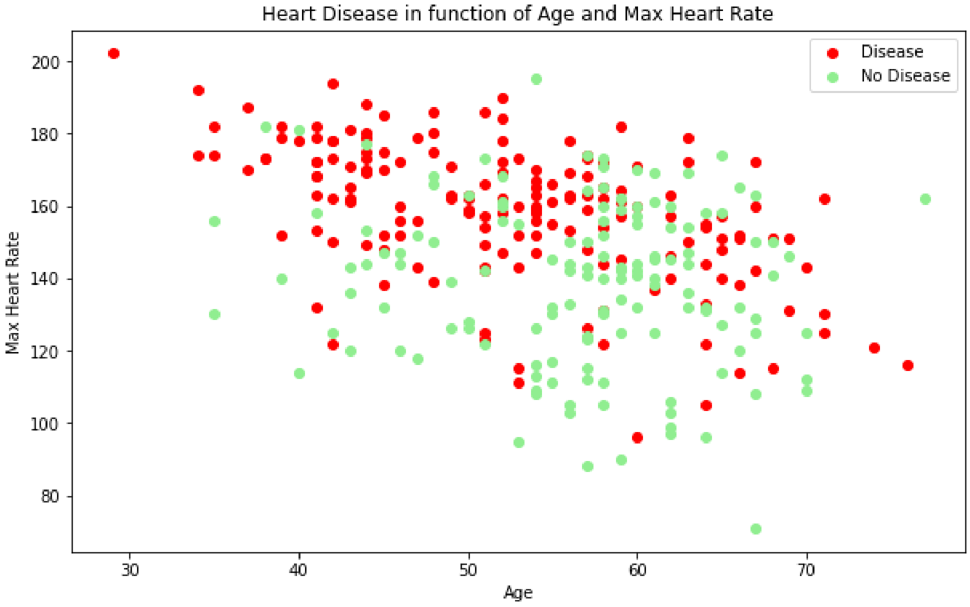


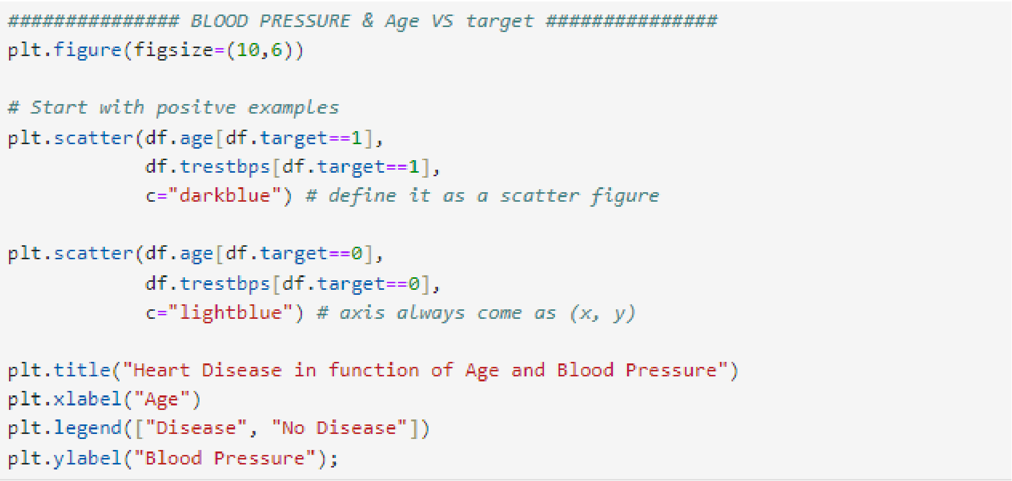
```
###################### HEART RATE (thalach) & AGE VS TARGET ################
# Create another figure
plt.figure(figsize=(10,6))

# Start with positve examples
plt.scatter(df.age[df.target==1],
            df.thalach[df.target==1],
            c="red") # define it as a scatter figure

plt.scatter(df.age[df.target==0],
            df.thalach[df.target==0],
            c="lightgreen") # axis always come as (x, y)

plt.title("Heart Disease in function of Age and Max Heart Rate")
plt.xlabel("Age")
plt.legend(["Disease", "No Disease"])
plt.ylabel("Max Heart Rate");
```



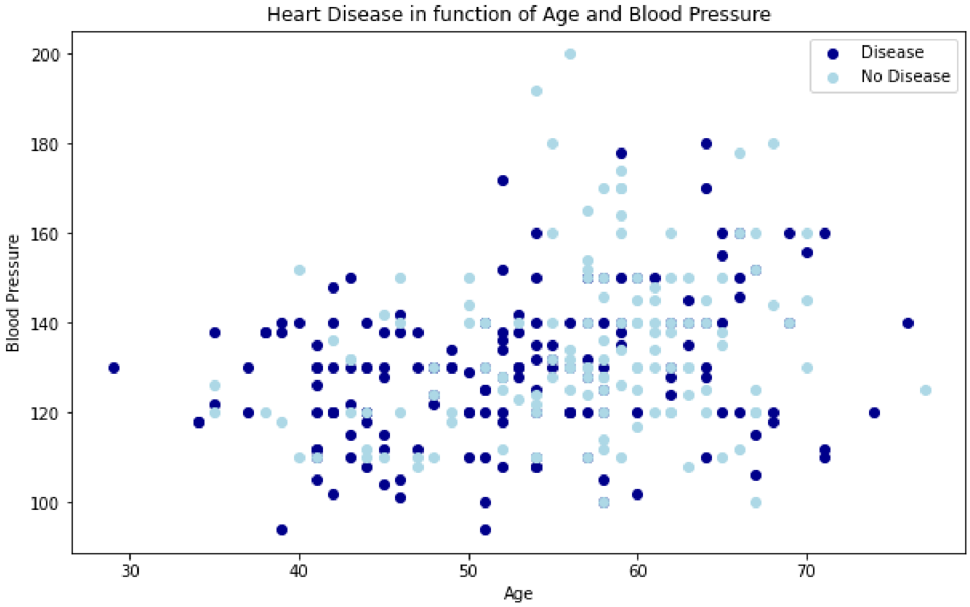


```
############### BLOOD PRESSURE & Age VS target ###############
plt.figure(figsize=(10,6))

# Start with positve examples
plt.scatter(df.age[df.target==1],
            df.trestbps[df.target==1],
            c="darkblue") # define it as a scatter figure

plt.scatter(df.age[df.target==0],
            df.trestbps[df.target==0],
            c="lightblue") # axis always come as (x, y)

plt.title("Heart Disease in function of Age and Blood Pressure")
plt.xlabel("Age")
plt.legend(["Disease", "No Disease"])
plt.ylabel("Blood Pressure");
```



- **Conclusion:** In conclusion, this comprehensive case study demonstrates the utility of machine learning in predicting heart disease risk and guiding preventive interventions in clinical practice. By leveraging advanced analytics and patient data, machine learning algorithms enable more accurate and timely identification of individuals at heightened risk of heart disease, ultimately contributing to better health outcomes and reduced disease burden. Moving forward, continued research and innovation in machine learning and healthcare analytics hold the potential to further enhance cardiovascular risk assessment and management, ushering in a new era of precision medicine in cardiology.