HEART DISEASE PREDICTION

Overview:

This is a project on heart disease prediction dataset. We will perform data analysis then we shall build models such as logistic regression and decision trees by splitting data into training and test sets. The essence of this process is obtain predictions from our models.

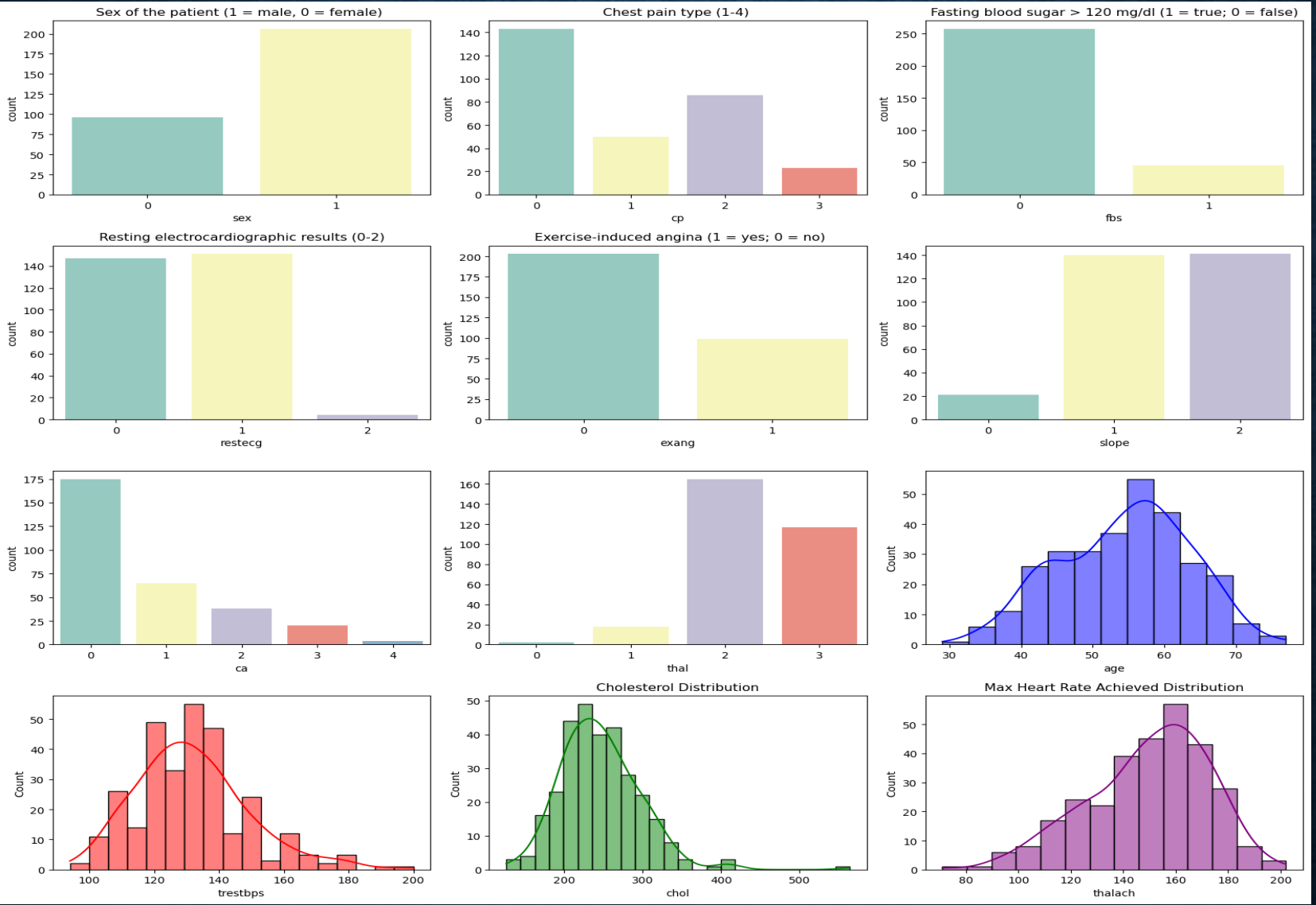
Business and Data Understanding

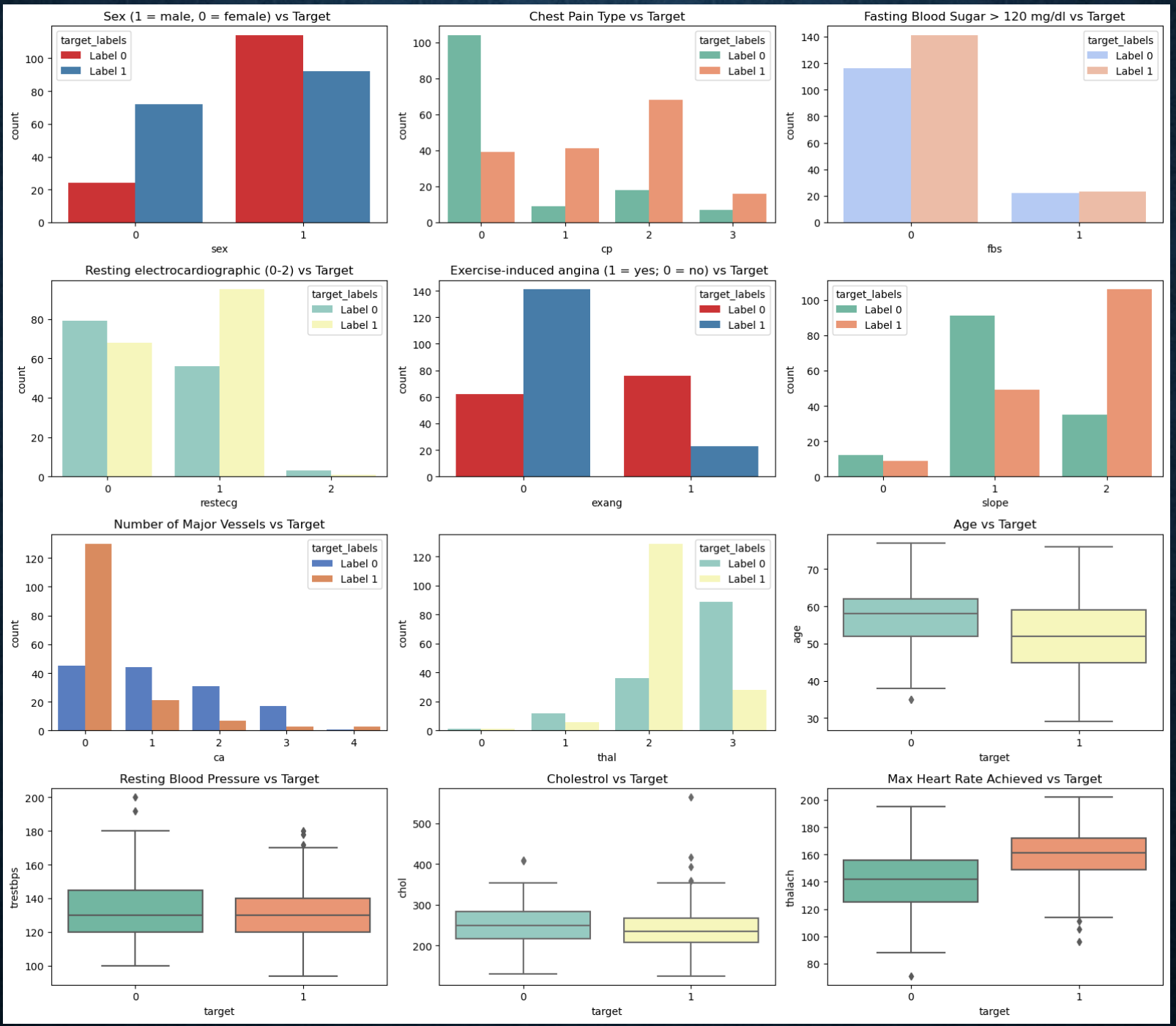
- I have chosen a heart disease prediction dataset, my stakeholder is a hospital administrator. We would like to predict the likelihood of a patient having a heart disease based on their medical history and other factors. In order to help the hospital administrator make informed decisions about resource allocation and patient care.
- A classification problem in data science is a type of supervised learning problem where the goal is to predict a categorical label or class that an instance belongs to, based on its features or characteristics.

Benefits of using classification

1. **Early Intervention and Prevention**
2. Improved Diagnosis and Treatment
3. Enhanced Patient Outcomes
4. Reduced Healthcare Costs
5. Personalized Medicine

Modeling


Sex of the patient (1 = male, 0 = female)
count
sex
Chest pain type (1-4)
cp
Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
fbs
Resting electrocardiographic results (0-2)
restecg
Exercise-induced angina (1 = yes; 0 = no)
exang
slope
ca
thal
Count
age
Cholesterol Distribution
trestbps
chol
Max Heart Rate Achieved Distribution
thalach


Sex (1 = male, 0 = female) vs Target
target_labels
Label 0
Label 1
count
sex
Chest Pain Type vs Target
cp
Fasting Blood Sugar > 120 mg/dl vs Target
fbs
Resting electrocardiographic (0-2) vs Target
restecg
Exercise-induced angina (1 = yes; 0 = no) vs Target
exang
slope
Number of Major Vessels vs Target
ca
thal
Age vs Target
age
target
Resting Blood Pressure vs Target
trestbps
Cholestrol vs Target
chol
Max Heart Rate Achieved vs Target
thalach

Evaluation

Based on our logistic regression we had specific metric such as:

Accuracy of 1.00

Precision: 1.00

Recall: 1.00

F 1 Score: 1.00

Confusion Matrix: [[29   0]

[0   32]]

Evaluation

Based on the Decision Trees we had the metrics:

Accuracy: 0.77

Precision: 0.71

Recall: 0.86

F1-score: 0.78

Confusion Matrix: [[25 4]

[10 22]]

Recommendations

1. **Collect more data**: Increasing the size of the dataset can lead to more accurate models. Consider collecting data from multiple sources, such as hospitals, clinics, or wearable devices.
2. **Feature engineering**: Extract more relevant features from the existing data, such as calculating the body mass index (BMI) from height and weight, or extracting relevant information from medical histories.
3. **Hyperparameter tuning**: Perform hyperparameter tuning for the logistic regression and decision tree models to optimize their performance.
4. **Ensemble methods**: Consider using ensemble methods, such as bagging or boosting, to combine the predictions of multiple models and improve overall accuracy.
5. **Deep learning models**: Explore the use of deep learning models, such as neural networks, to capture complex relationships in the data

Next Steps

- Explore hyperparameter tuning for both models to improve performance.
- Investigate feature engineering techniques to improve model accuracy.
- Consider ensemble methods to combine the strengths of both models.

THANK YOU