HEART DISEASE PREDICTION USING DIFFERENT CLASSIFIERS

Arzoo Kharayat

6th October, 2022

Abstract
Heart disease describes a range of conditions that affect the heart such as diseased vessels, structural problems and blood clots. This report has analyzed different classification models used for predicting the presence of heart disease. The dataset which has been studied consists of 14 attributes - 13 labels and 1 target. It consists of 1025 entries/rows. The classification models trained on the data are - Logistic regression, KNN, SVM, Decision trees and Naive Bayes and their accuracy scores are analyzed.

1. Problem Statement

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. In order to prevent this staggering amount of fatalities, it is crucial to identify those at highest risk of CVDs and ensure that they receive appropriate and timely treatment. It can be useful to develop an accurate and robust system which can determine whether an individual is likely to develop heart disease. This can be done by gathering and analyzing data, making meaningful interpretations and then designing algorithms to predict new data which is why Machine learning proves to be a groundbreaking success in this sphere.

2. Market / Customer / Business Need Assessment

In order to diagnose a patient with heart disease, the healthcare provider examines personal and family medical history. After that, a series of tests is carried out. Besides blood tests and chest X-rays, tests to diagnose heart disease can include: Electrocardiogram (ECG or EKG), Holter monitoring, Echocardiogram, Exercise tests or stress tests, Cardiac catheterization, Heart (cardiac) CT scan and Heart (cardiac) magnetic resonance imaging (MRI) scan. These tests fall under the category of non-invasive tests (which do not involve tools that break the skin or physically enter the body). Sometimes non-invasive tests don't provide enough answers which calls for the need to use an invasive procedure to diagnose heart disease.
After collecting this data, the decision is usually made by the doctors on the basis of their intuition and experience.This practice can be problematic as the doctors can come to an incorrect conclusion due to errors, unwanted bias or just a lack of knowledge.
The objective of the system is to provide a uniform framework that can be used worldwide to predict the presence of heart disease and the various models are analyzed to obtain the best accuracy. This system can also prove to be useful by helping non-specialized doctors diagnose heart disease.

3. Target Specifications and Characterization

This system is designed to bring about improvement in the diagnosis and prediction of heart disease. It is therefore, targetted towards the general population to gauge their cardiovascular health. It can be used by healthcare providers to gain a more thorough understanding of the patient's symtpoms and can aid them in diagnosis and further treatment.
The use of this system coupled with domain knowledge and experience can drastically reduce the fatalities caused due to cardiovascular ailments.

4. External Search

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. It can be found on UCI Machine Learning Repository.

The dataset consists of 14 attributes.
Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. target:0 for no presence of heart disease, 1 for presence of heart disease

References:
Raschka, Sebastian, & Mirjalili, Vahid. (2019).*Python Machine Learning*

Links:
https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
https://www.healthline.com
https://www.geeksforgeeks.org
Link for the dataset: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

5. Benchmarking alternate products

Xinjiang is located in northwest China and is home to multiple ethnic groups. Uyghur and Kazakh are the main ethnic groups in Xinjiang. Studies found that these populations have high prevalence of CVD risk factors, such as metabolic syndrome, hypertension, and obesity, thereby corresponding with high incidence of CVD. There was a study aimed to use ML algorithms to establish a CVD that was suitable for the Xinjiang Uyghur and Kazak Populations based on routine physical examination indicators.

The research cohort data collection was divided into two stages. The first stage involved a baseline survey from 2010 to 2012, with follow-up ending in December 2017. The second-phase baseline survey was conducted from September to December 2016, and follow-up ended in August 2021. A total of 12,692 participants (10,407 Uyghur and 2,285 Kazak) were included in the study.

After 4.94 years of follow-up, a total of 1,176 people were diagnosed with CVD (cumulative incidence: 9.27%). Therefore, we can see how effective prediction systems can help save lives.

6. Applicable Patents

No patents required as all the software and the algorithms used are open source.

7. Applicable Regulations

No relevant applicable regulations found.

8. Applicable Constraints

We will need to have a better understanding of the domain to further improve the state of the models. This can be done by consulting the experts in the field of medicine such as researchers and doctors.

We will need better GPUs to preprocess large amounts of data and visualize it so our model has more data to train on and be ready for all the problems found in the real world and ensure that no patient goes unnoticed.

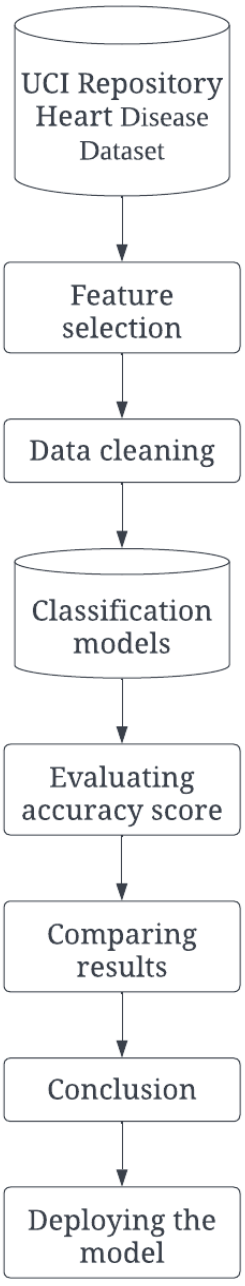
9. Business Model

As there are no similar finished products in the market we can look towards monopolizing the market and sell the finished product to the government at subsidized rates so that it can be used in government hospitals for greater good. We can also expand this in the private sector at profitable rates and that funding can further aid in ongoing research to better our algorithm with future updates.

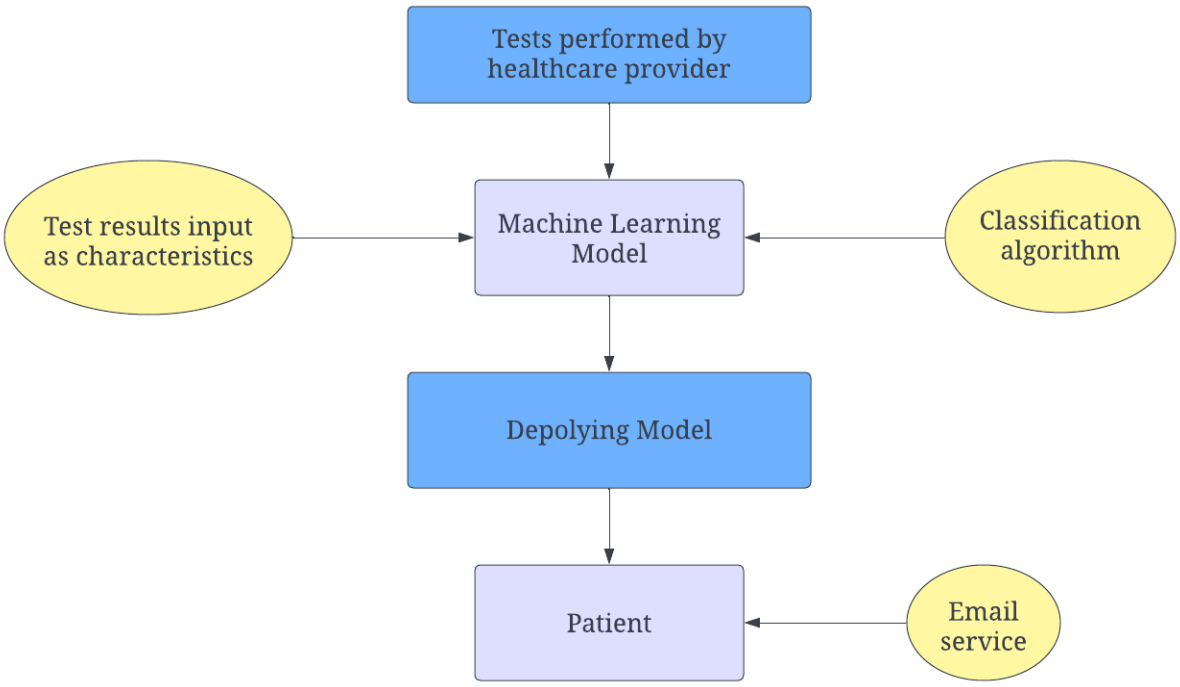
10. Concept Generation

Health issues have been on the rise in the past decade and misdiagnosis has been one of the major reasons behind it. AI researchers recognised this and the idea of using AI as a means of diagnosis was born to eliminate the factor of human error while diagnosing , the idea isn't to replace doctors who are diagnosing the patients but to give them a guiding line which gives them a better idea to diagnose the patient correctly. It also gives a patient a sense of where they stand as they can always refer to the numbers outputted by the algorithm rather than having no option but to blindly rely on what the doctor says. AI Diagnosis is going to be one of the most important fields going forward and it'll improve our health sector for good.

11. Concept Development



12. Final Product Prototype



13. Product Details

13.1. How does it work?

We implement different classification models on the dataset to figure out the best possible algorithm for a real world scenario.

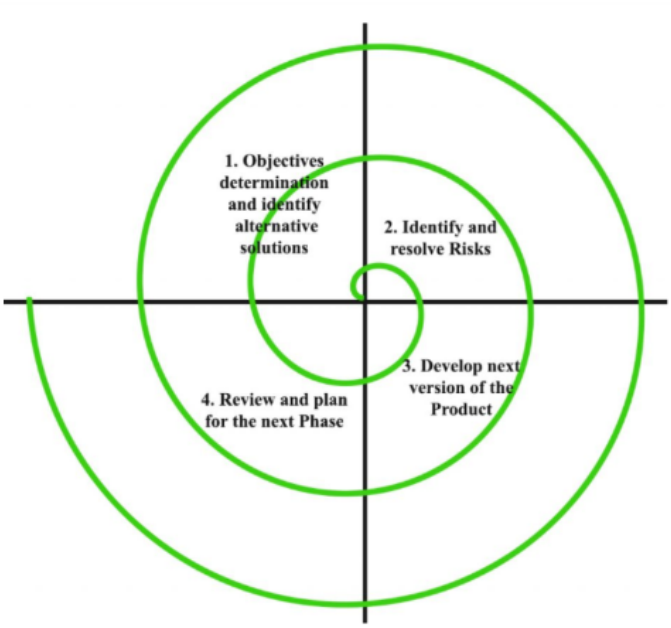
Software engineering is the process of applying science, tools and methods to find cost effective solutions to problems. It is the systematic, disciplined and quantifiable approach for the development, operation and maintenance of software.

There are many models available that are used to develop software like- Iterative Waterfall model, Incremental model, RAD model, Spiral model, Agile model etc. The most appropriate model for this system is Spiral model.

Phases of spiral model:

1. Requirement analysis – This phase includes requirement gathering and analysis. Based on the requirements, objectives are defined and different alternate solutions are proposed.
2. Risk analysis and resolving – In this quadrant, all the proposed solutions are analyzed and any

potential risk is identified, analyzed, and resolved.
3. Develop and test**:** This phase includes the actual implementation of the different features. All the implemented features are then verified with thorough testing.
4. Review and planning of the next phase **:** In this phase, the software is evaluated by the customer. It also includes risk identification and monitoring like cost overrun or schedule slippage and after that planning of the next phase is started.



13.2. Data sources

The dataset has been downloaded from the UCI machine learning repository.

13.3. Algorithms

The algorithms used are as follows:

13.3.1. K-Nearest Neighbors

The KNN algorithm is fairly straightforward and can be summarized by the following steps:

- Choose the number *k* and a distance metric.
- Find the *k*-nearest neighbors of the data record that we want to classify.
- Assign the class label by majority vote.

Based on the chosen distance metric, the KNN algorithm finds the *k* examples in the training dataset that are closest (most similar) to the point that we want to classify. The class label of the data point is then determined by a majority vote among its *k* nearest neighbors.

13.3.2. Logistic Regression

Logistic regression is a classification model that is very easy to implement and performs very well on linearly separable classes. It is widely used as a probabilistic model for binary classification and uses the concept of odds.

Odds can be written as $p / (1-p)$ where p stands for the probability of the positive event.

We then further define the *logit* function, which is simply the logarithm of the odds (log-odds):

$logit (p) = log (p / (1-p))$

To predict the probability of belonging to a certain class,we use the *logistic sigmoid function* which is the inverse of the logit function. It is called sigmoid because of it's characteristic S-shape.

13.3.3. Decision Trees

The Decision Tree model breaks down the data by making a decision based on asking a series of questions. Based on the features in our training dataset, the decision tree model learns a series of questions to infer the class labels of the examples.

Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain (IG). In an iterative process, we can then repeat this splitting procedure at each node until the leaves are pure.

This can result in a deep tree with many nodes which can easily lead to overfitting. Therefore, we want to prune the tree by setting a limit for the maximal depth of the tree.

13.3.4. Support Vector Machines

SVMs can be considered as an extension of the perceptron. In SVMs, our optimization objective is to maximize the margin. The margin is defined as the distance between the separating hyperplane(decision boundary) and the training examples that are the closest to this hyperplane, which are the so-called support vectors.

13.3.5. Naive Baiyes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. It is called Bayes because it depends on the principle of Bayes Theorem.

There are three types of Naive Bayes Model, which are given below:

- **Gaussian**: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- **Multinomial**: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.
  The classifier uses the frequency of words for the predictors.
- **Bernoulli**: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

13.4. Team

A team of 5-8 people is required for ANN implementation to achieve accurate real world results. The following roles will be required:

- Data engineer
- Machine learning engineer
- Cloud engineer
- Software Developer
- Researcher / Domain expert

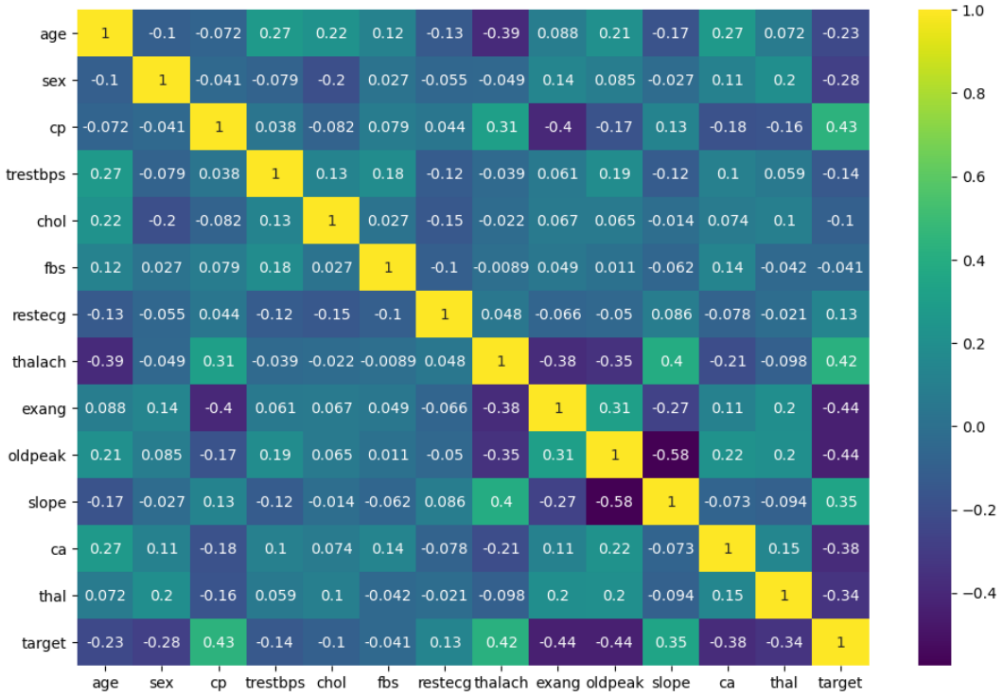
14. Code Implementation

Link to the github repository: https://github.com/hell0worId/hd-prediction

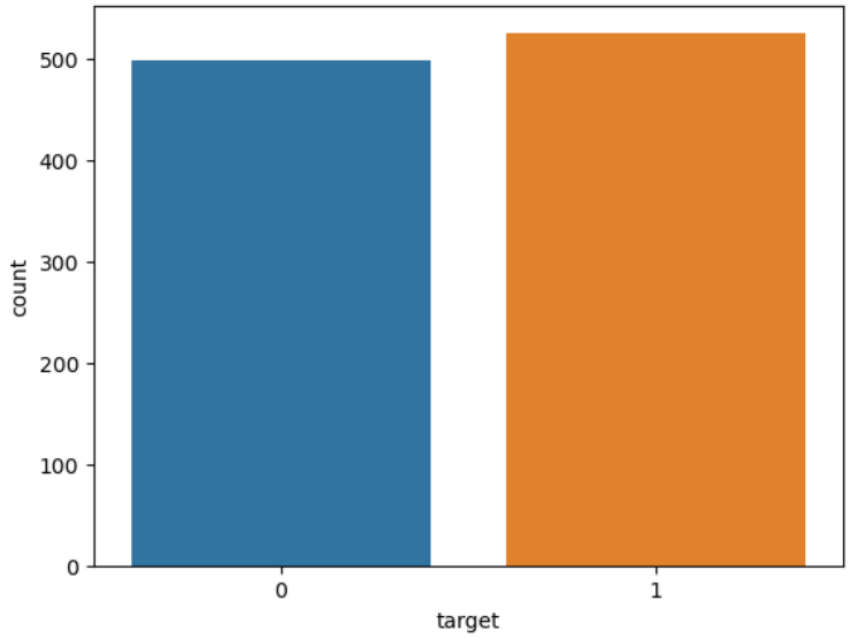
Data visualization:

In [8]:
```
plt.figure(figsize=(12,8))
sns.heatmap(df.corr(),annot=True,cmap="viridis")
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x12d2fed31c0>

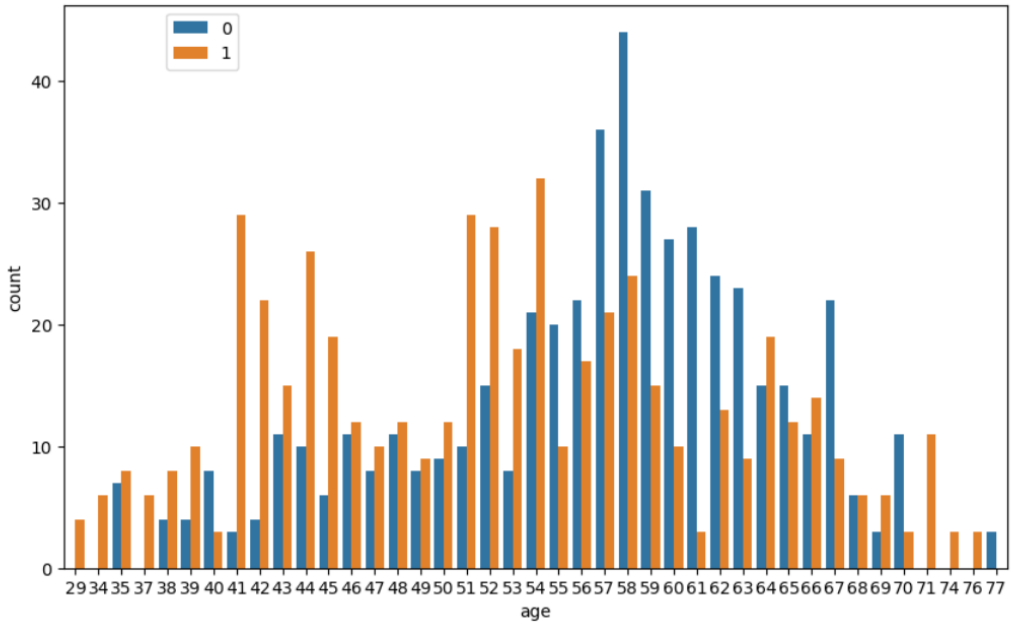


In [9]:
```
#Value counts of each target
sns.countplot(data=df,x="target")
plt.show()
```



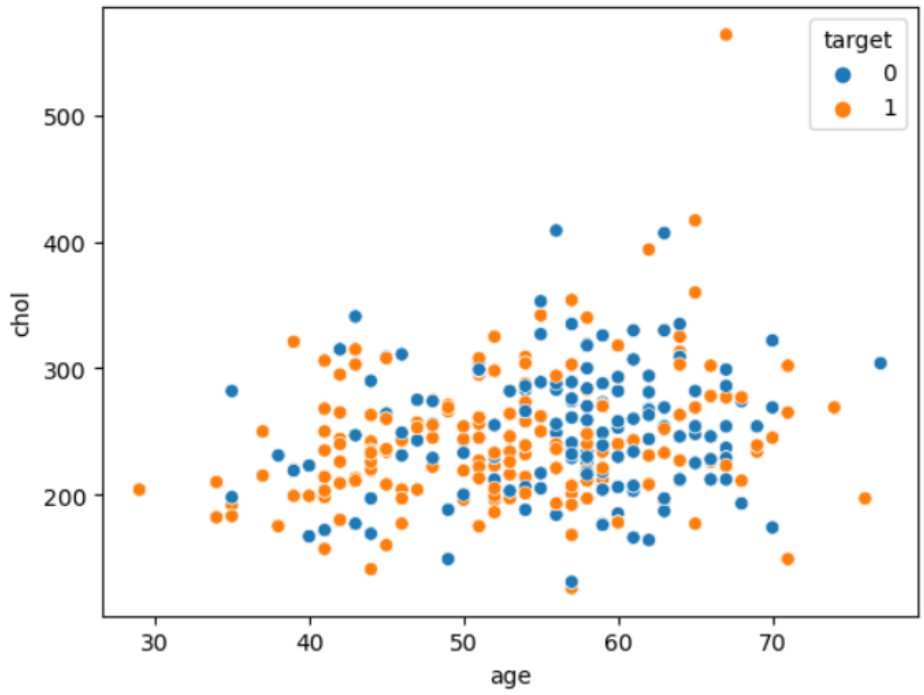
In [10]:

```
plt.figure(figsize=(10,6))
sns.countplot(data=df,x="age",hue="target")
plt.legend(bbox_to_anchor=(0.10,1))
plt.show()
```



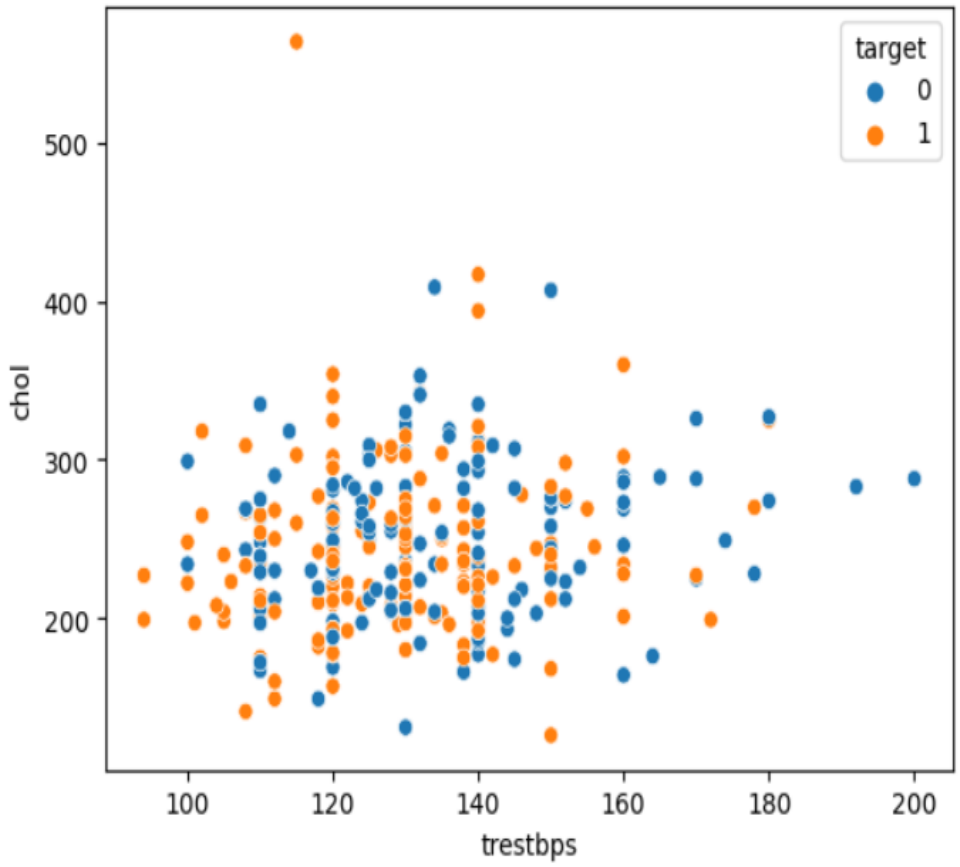
In [12]:

```
sns.scatterplot(data=df,x="age",y="chol",hue="target")
plt.show()
```



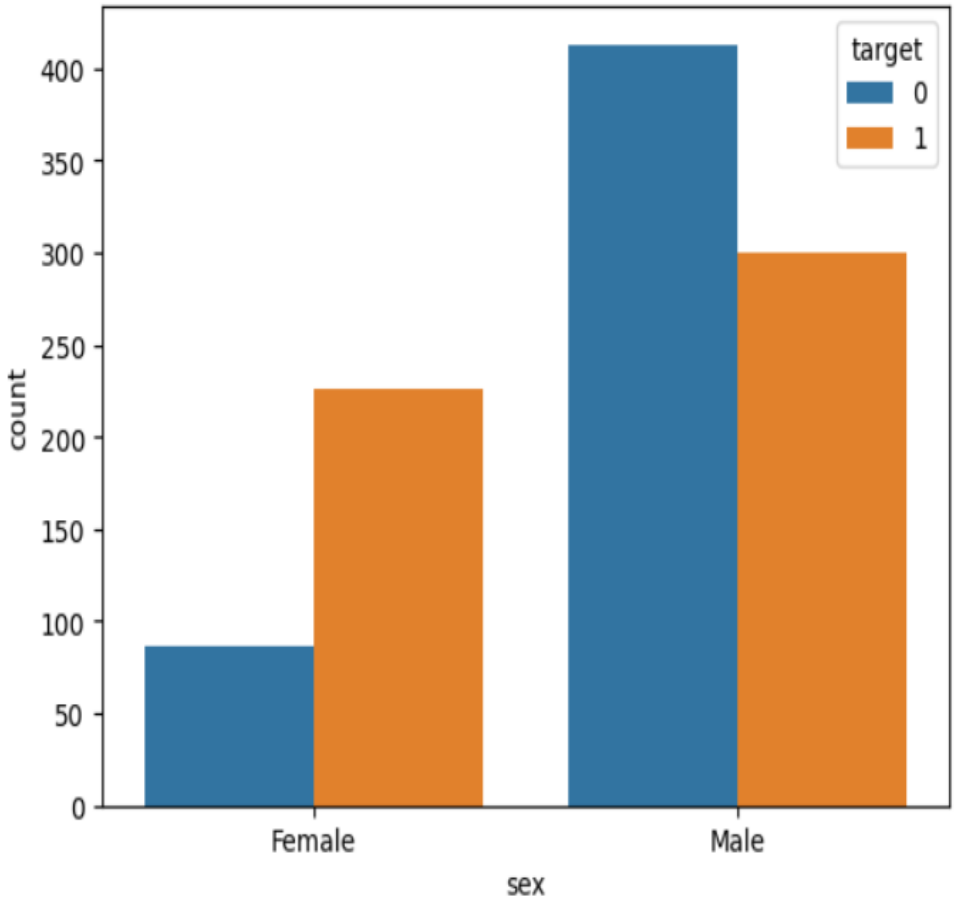
In [13]:

```
sns.scatterplot(data=df,x="trestbps",y="chol",hue="target")
plt.show()
```



In [14]:

```
g = sns.countplot(data=df,x="sex",hue="target")
g.set_xticklabels(["Female","Male"])
plt.show()
```



15. Conclusion

The various classification algorithms used were- Logistic Regression, KNN, SVM, Naive Bayes and Decision Trees. After training the model on each classifier, the accuracy scores were computed and recorded. Based on the observations, Decision Trees has the best accuracy and is therefore, most viable for a heart disease prediction system.