San José State University
Computer Science Department

CS156, Introduction to Artificial Intelligence, Fall 2021

Term project progress report

Name: Hanh Nguyen

SID: 014876069

Provide the following information for your term project.

1. **AI problem and background:** Describe the problem. What is the prediction problem? Why is it interesting? Why is it important to solve? Have there been other previous solutions? What are they like? This section also includes any information that is relevant to the reader's understanding of the problem. Assume that whoever is reading this description has no knowledge of the problem at hand. Explain clearly and concisely what they need to know in order to understand what problem you are trying to solve and why. What has been previously done to solve this problem? Have others (companies or individuals, or maybe you) attempt to arrive at a solution for this problem? Do you agree with their approach? Anything you would do differently?

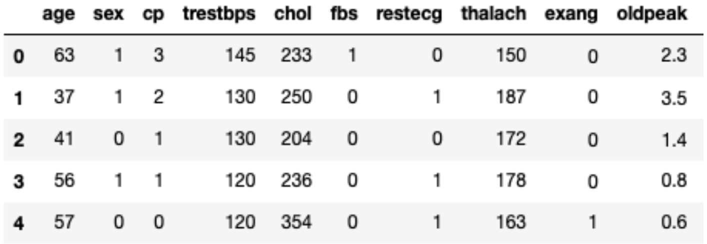
   Prediction problem: "Predicting heart disease using machine learning." It is interesting because by using parameters in clinical on a patient, we can predict whether they can have heart disease or not have heart disease. One of the most dangerous diseases is cardiovascular disease. Heart disease is taking the lives of many people all over the world. There are many causes related to cardiovascular diseases, such as diabetes and hypertension. These diseases need early detection to prevent heart diseases. Therefore, model machine learning to predict heart diseases will help in this situation. There have been other previous solutions, such as using "strength scores with significant predictors." Although there are many predictable solutions for heart diseases, the final result is diagnosing and predicting heart diseases based on current clinical tests of patients. So, I agree with the approach I referenced on the link: https://www.kaggle.com/faressayah/predicting-heart-disease-using-machine-learning/notebook. There are a few things I will do differently. I will use a pair plot in the seaborn library to plot statistics, a decision tree classifier to use sub-features in different classification stages, and normalization databases. The purpose of normalization is to store logical data and remove useless data.

2. **Dataset of choice:** Describe the dataset you are going to use to solve the above described problem. Provide the source for this dataset. Confirm that you were already able to download this dataset.

   I am going to use "Heart Disease UCI" datasets from the URL: https://www.kaggle.com/ronitf/heart-disease-uci. I am done with downloading the Heart.csv dataset.

3. **Describe the independent variables:** Describe the data fields you are going to use to train your model. Describe all the independent variables. What do the columns of your dataset represent? What kinds of variables are they (e.g. numeric vs. categorical). You might want to include here a table describing all the independent variables. Or maybe your training data consists of free text (NLP problems)? Describe what your training data represents. How many variables are you using (how big is your feature space)?

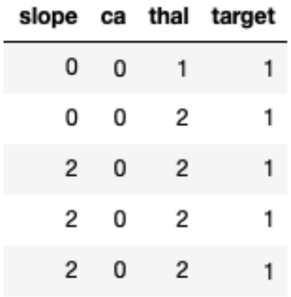
   The independent variable: age, sex, chest pain type(cp), resting blood pressure(trestbps), serum cholesterol(chol), fasting blood sugar(fbs), resting electrocardiographic(restecg), maximum heart rate achieved (thalach), exercise-induced angina(exang), ST depression induced by exercise relative to rest(oldpeak). The columns of the dataset represent all the information about the patient, such as age, sex, and parameter in clinical. The variables are numeric.



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

4. **Describe the dependent variable:** Describe the dependent/response variable. What kind of variable is it (e.g. numeric vs. categorical)? Is this variable a proxy for something your model is aiming to predict? Does this variable suggest the type of problem you are solving (e.g. regression vs. classification)? If dealing with a classification problem, state the number of classes in the dataset.

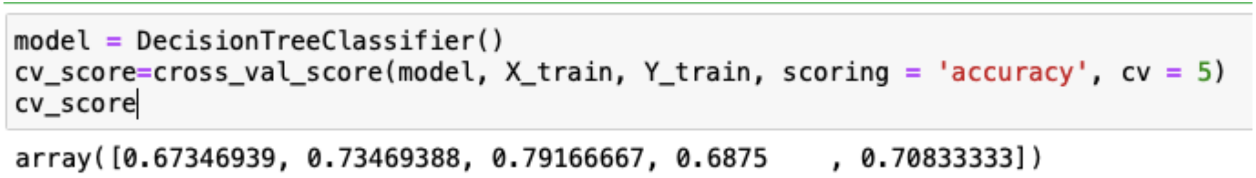
   Four dependent variables on the right are the output of the process. They are also numeric variables.



slope	ca	thal	target
0	0	1	1
0	0	2	1
2	0	2	1
2	0	2	1
2	0	2	1

5. **Describe the data splits:** Describe here also your training vs. validation vs. test split. Or maybe you are going to use cross-validation approach and do not have a set aside validation dataset. Did you stratify your split?

   For example: using a cross-validation approach

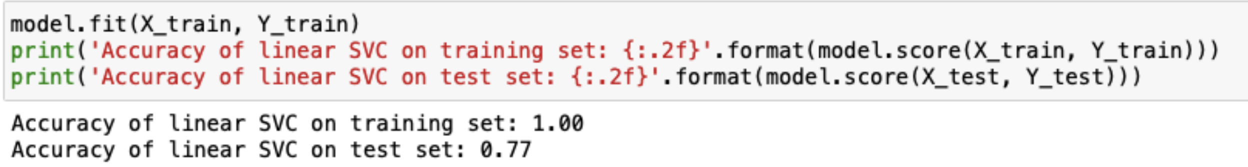


```
model = DecisionTreeClassifier()
cv_score=cross_val_score(model, X_train, Y_train, scoring = 'accuracy', cv = 5)
cv_score
```

```
array([0.67346939, 0.73469388, 0.79166667, 0.6875    , 0.70833333])
```

6. **Number of training observations:** State the number of observations in your training data. If dealing with a classification problem, list the number of observations in each class.

   Accuracy of linear SVC on training set

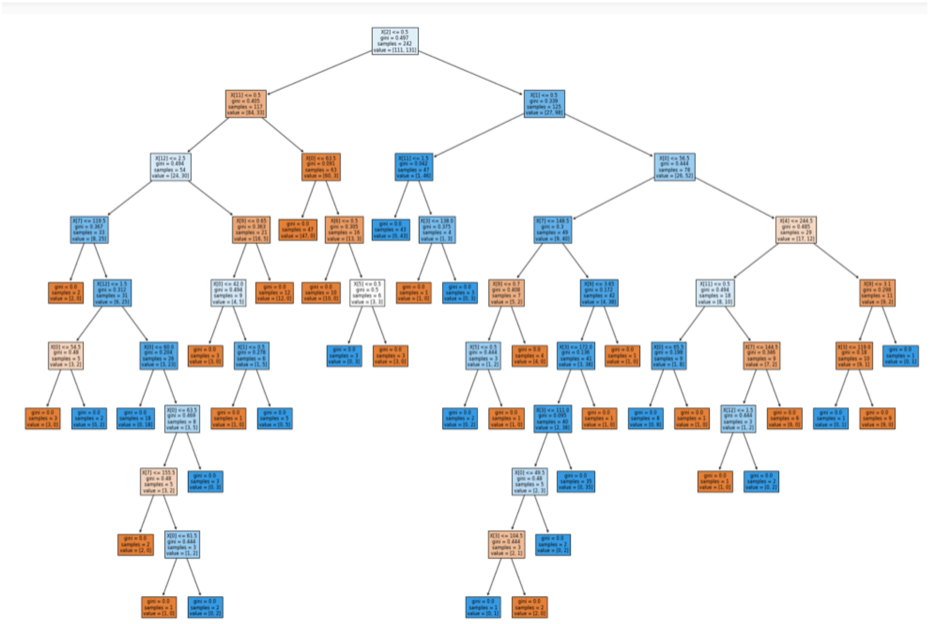


```
model.fit(X_train, Y_train)
print('Accuracy of linear SVC on training set: {:.2f}'.format(model.score(X_train, Y_train)))
print('Accuracy of linear SVC on test set: {:.2f}'.format(model.score(X_test, Y_test)))
```

```
Accuracy of linear SVC on training set: 1.00
Accuracy of linear SVC on test set: 0.77
```

Model training:

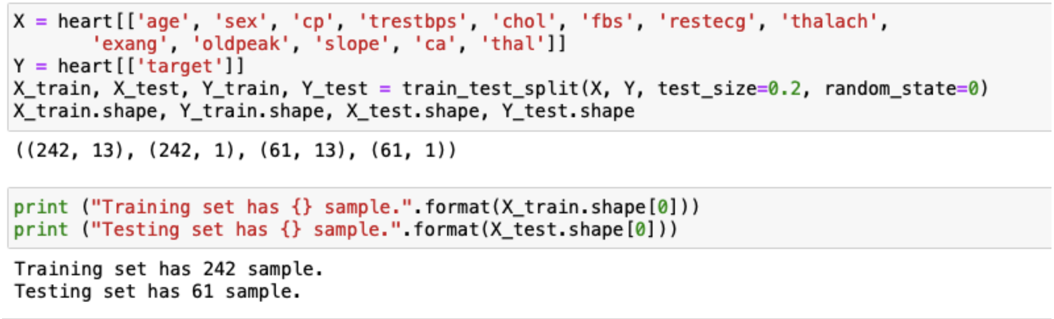
The goal is to create a model that predicts values using simple rules.



7. **Number of validation observations:** State the number of observations in your validation data (if using a set aside validation dataset). Otherwise, just put N/A here.
   N/A

8. **Number of test observations:** State the number of observations in your test data. If dealing with a classification problem, list the number of observations in each class.

   Training and testing data sets



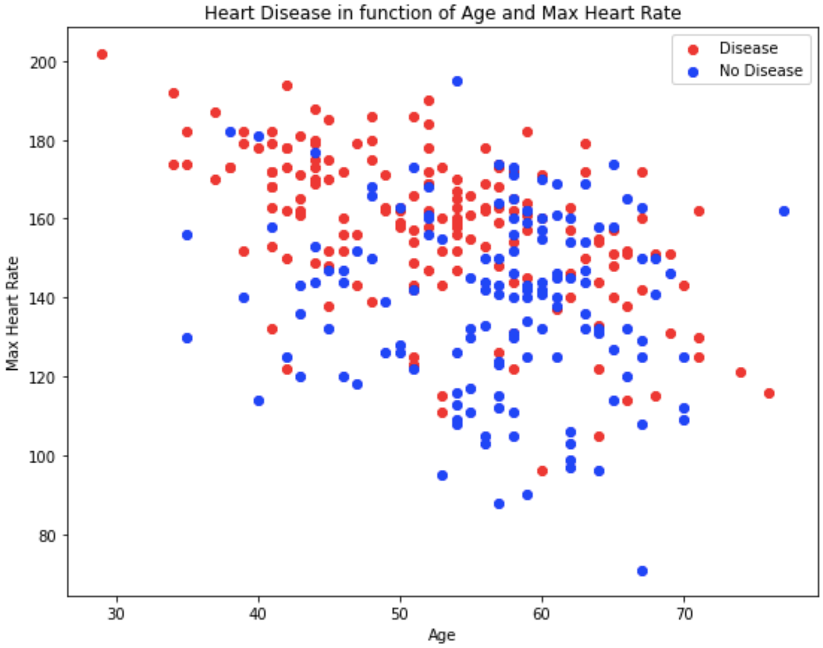
```
X = heart[['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal']]
Y = heart[['target']]
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=0)
X_train.shape, Y_train.shape, X_test.shape, Y_test.shape
```

```
((242, 13), (242, 1), (61, 13), (61, 1))
```

```
print ("Training set has {} sample.".format(X_train.shape[0]))
print ("Testing set has {} sample.".format(X_test.shape[0]))
```

```
Training set has 242 sample.
Testing set has 61 sample.
```

Heart Disease of Age and Max Heart Rate



Decision tree in ML can be used for both classification and regression supervised problems.