Janet Anagli and Christoph F. Eick

COSC 3337: Data Science I in Fall 2024

ProblemSet1 Task2

 Individual Task [[1]](#footnote-1)

**Task 2: Predicting Heart Disease using Decision Trees and Support Vector Machines**

****

**Submission deadline:** Sept. 27, 11:59p CST (electronic submission)

**Weight:** 13% of the points that are allocated to the Problem Set Tasks

Responsible TA: Janet

Last updated: Sept. 13, 2 pm.

**Learning Objectives**:

1. Understand how decision trees and support vector machines (SVMs) are used for classification tasks.
2. Learn to tune hyperparameters in decision trees and SVMs to avoid overfitting.
3. Gain proficiency in using recall, precision, and accuracy as metrics for evaluating classification model performance.
4. Learn to use cross-validation techniques to assess model performance.

In this assignment, you will work with a dataset containing medical data used for predicting heart disease. Your goal is to build, evaluate and compare two models for predicting whether a patient has a heart disease or not: one using Decision Tree and one using SVM.

**Overview:**
Heart disease refers to a range of conditions that affect the heart, often linked to blockages in the heart's arteries or irregular heart function. Risk factors include age, gender, high cholesterol, high blood pressure, and lifestyle factors such as exercise habits. The dataset provided includes various medical attributes that help in predicting heart disease. These factors, such as chest pain type, cholesterol levels, blood pressure, and exercise responses, are crucial in assessing an individual's risk. By analyzing these variables, machine learning models like decision trees and support vector machines (SVMs) can help predict the likelihood of heart disease and support preventive measures.

**Dataset:**

*Optional*  **Download the “heart-disease2.csv” dataset** [**here**](https://uofh-my.sharepoint.com/%3Ax%3A/g/personal/jyanagli_cougarnet_uh_edu/EfBsGgyHzN5AuWxvvHXh0SsBXkfQBoMWy2mvVn5RivYm_w?e=CTSwbe)**.**

This dataset contains medical data used for predicting heart disease. The data includes various attributes such as age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), and ST depression induced by exercise relative to rest (oldpeak). Attributes “chest pain type (cp)” and “resting electrocardiographic results (restecg)” have been one-hot encoded because those attributes are categorical, not nominal-the distinction is especially essential for SVMs.

The data was collected from patients during regular medical check-ups and heart disease screenings. Standard medical procedures and equipment were used to record the various attributes. Data was anonymized to protect patient privacy.

More details about the attributes of this dataset are listed below:

1. age: Age of the patient (in years) - Domain [29, 77],
2. sex: Sex of the patient (1 = male, 0 = female),
3. trestbps: Resting blood pressure (in mm Hg on admission to the hospital) – Domain [94, 200],
4. chol: Serum cholesterol in mg/dl – Domain [126, 564],
5. fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false),
6. thalach: Maximum heart rate achieved – Domain [71, 202],
7. exang: Exercise-induced angina (1 = yes; 0 = no),
8. oldpeak: ST depression induced by exercise relative to rest – Domain [0, 6.2],
9. slope: Domain [0, 2],
10. ca: Domain [0, 4],
11. thal: Domain [0, 3],
12. cp: Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic),
13. restecg: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy) and
14. target: (Heart Disease Status: 1 = presence of heart disease, 0 = absence of heart disease; class attribute that needs to be predicted using SVM and Decision Trees).

The first 3 examples of the dataset are listed below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | sex | trestbps | chol | fbs | thalach | exang | oldpeak | slope |
| 63 | 1 | 145 | 233 | 1 | 150 | 0 | 2.3 | 0 |
| 37 | 1 | 130 | 250 | 0 | 187 | 0 | 3.5 | 0 |
| 41 | 0 | 130 | 204 | 0 | 172 | 0 | 1.4 | 2 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ca | thal | cp\_0 | cp\_1 | cp\_2 | cp\_3 | restecg\_0 | restecg\_1 | restecg\_2 | target |
| 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 0 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 0 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

**Tasks**

1. A. Using all attributes, **build a Decision Tree model** to predict whether a patient has a heart disease or not: Train the Decision Tree model using the given maximum depths: 3, 7, 11, 15.  **8 points**
2. Perform 5-fold cross-validation for each of the 4 max depths and compute accuracy (mean of validation scores), precision and recall. Generate a table, as given below, for the obtained results. **5 points**

|  |
| --- |
| **Decision Tree Experiments** |
| Max Depths | Accuracy | Precision | Recall |
| 3 |  |  |  |
| 7 |  |  |  |
| 11 |  |  |  |
| 15 |  |  |  |

1. Explain how the tree size/depth affects model performance in the context of overfitting/underfitting. **3 points**

1. Explain the meaning of the difference in accuracy, precision and recall scores in relation to the task; only if there is a significant difference. **2 points**
2. A. Using all attributes, **build an SVM Model** to predict whether a patient has a heart disease or not: Train the SVM model using the given **4 kernel functions**: linear, polynomial, sigmoid, sigmoid with different σ value. **10points**

**Note:**

Excluding the 4th kernel function where you use a different σ ,“coef0” value (“coef0” is the σ equivalent using Python scikit-learn-By default, coef0 is set to 0.0. Use the σ equivalent in your chosen programming language.) for the sigmoid kernel, use the default values for the parameters of the other **3 SVM kernel functions** (linear, polynomial, sigmoid).

Adjust the σ parameter of the sigmoid kernel to fine-tune the SVM model for better model performance, this represents the 4th kernel function where you’re supposed to use sigmoid with different σ value.

B. Perform 5-fold cross-validation for each of the 4 kernel functions and compute accuracy (mean of validation scores), precision and recall. Generate a table, as given below, for the obtained results. **5 points**

|  |
| --- |
| **SVM Experiments** |
| **Kernel Function** | Accuracy | Precision | Recall |
| linear |  |  |  |
| polynomial |  |  |  |
| sigmoid |  |  |  |
| sigmoid with different σ value |  |  |  |

1. Discuss the impact of different kernels on model performance. **2 points**
2. Explain the meaning of the difference in accuracy, precision and recall scores in relation to the task. **2 points**
3. Interpret the tables you generated in questions 1B and 2B; compare the performance of the Decision Tree and SVM models. Which model performs better? Why do you think that is the case? What would you recommend to further improve each model’s performance? **5 points (and up to 3 extra points)**

## **Deliverables:**

1. A python/R file with your code and analysis.
2. A pdf copy of your word document report (1-2 pages) of your findings and conclusions for Task 2C, 2D, 3C, 3D and 4. Paste the specific code and the respective results into your report before adding your conclusions. Format the task name with “Bold” before explanations.

Note: If you use a Python notebook like Jupyter Notebook, you can include your conclusions in the notebook right after each question (no word document required), submit the python notebook file and submit a pdf copy of the notebook.

**Submission Guidelines[[2]](#footnote-2):**

1. Name your python/R files to **COSC3337F24-PS1T2-Firstname-Lastname.ipynb** or any other appropriate format.
2. Name the pdf copy of your report **COSC3337F24-PS1T2-Report-Firstname-Lastname.pdf**
3. Create a folder and name it **COSC3337F24-PS1T2-Firstname-Lastname**.The folder should contain both python/R file and pdf copy of your report named correctly. Compress (zip) the folder and submit it to MS TEAMS.
4. Submit on the Assignment tab in MS Teams.

**Guideline for Code Grading:**

1. Code documentation and readability: Use of comments, proper indentation, clear notations, and simplicity.
2. Code completeness: Working code with no errors including instructions for running your code. Indicate any module(s) to be downloaded before code is run.

**References:**

Python:

1. <https://www.datacamp.com/tutorial/decision-tree-classification-python>
2. <https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>

R:

1. <https://www.datacamp.com/tutorial/decision-trees-R>
2. <https://www.datacamp.com/tutorial/support-vector-machines-r>
1. Collaboration with other students is not allowed. [↑](#footnote-ref-1)
2. Updates for Task2 if any will be added to this specification by Sept. 23, 2024, the latest. [↑](#footnote-ref-2)