

Final Project

Machine Learning - Classification (COSC 22000)

25 Points

Group project: Maximum group members-2

Due Dates:

In-class presentation: April 18, 5.00 PM-7.00 PM (5 points)

Submission to DC Connect: April 20, 11.59 PM (20 points)

Mr. John Hughes has been reviewing the heartdata.csv dataset.

Let's Look at the Dataset.

Independent Variables

age - age in years

sex - (0 = female, 1 = male)

cp - chest pain type (0: typical angina, 1: atypical angina, 2: non-anginal pain, and 3: asymptomatic)

trestbps - resting blood pressure (in mm Hg on admission to the hospital)

chol - serum cholesterol in mg/dl

fbs - fasting blood sugar > 120 mg/dl - (0 = false, 1 = true)

restecg - resting electrocardiographic results – (0= Normal, 1= abnormality, 2= LVT)

thalach - maximum heart rate achieved

exang - exercise induced angina – (0 = no, 1 = yes)

oldpeak - ST depression induced by exercise relative to rest

slope - the slope of the peak exercise ST segment (0=up sloping, 1= flat, 2=down sloping)

ca - number of major vessels (0-4) colored by fluoroscopy

thal – feature is not defined – (0 = normal; 1=fixed, 2=reversible, 3=ND)

Dependent Variables

target – heart disease (0=no, 1=yes)

The Ask:

Build atleast **5** Learning models to predict the if a person has a heart disease or not.

One of those should be an Ensemble learning and one should be a Neural Network.

Below is the project template that you can use in your machine learning projects in Python.

1. Prepare Problem
 - a) Load libraries
 - b) Load dataset
2. Summarize Data/ Exploratory Data Analysis
 - a) Descriptive statistics
 - b) Data visualizations
3. Prepare Data
 - a) Data Cleaning
 - b) Feature Selection
 - c) Data Transforms
4. Model Building and Evaluate Algorithms
 - a) Split-out validation dataset
 - b) Test options and evaluation metric
 - c) Spot Check Algorithms
 - d) Compare Algorithms
5. Improve Performance
 - a) Algorithm Tuning
 - b) Ensembles
6. Finalize Model
 - a) Predictions on validation dataset
 - b) Create standalone model on entire training dataset
 - c) Save model for later use
7. Present your analysis of each of the above steps. Make conclusions and recommendations.
8. Use visualization when appropriate at any of the steps 1 to 7.
9. Use the best practices of model building we have been using discussing in our classes.
10. What will you do if the dataset was unlabeled. Assuming that the dataset is unlabeled, perform an unsupervised model building. Explain your findings.

Use `seed = 7` and `random_state= seed` when appropriate.

Note: Follow the best practices discussed throughout our classes.

Final Term Project Rubric

Scores	Exemplary (23-25)	Proficient (16-22)	Incomplete (11-15)	Needs Improvement (0-10)
(2%)	Clear description of the problem statement is given.	Mostly clear description of the problem statement is given.	An incomplete description of the problem statement is given.	Description of problem statement is incorrect or missing.
(5%)	The chosen EDA and data preprocessing are correct, meaningful and justified.	The chosen EDA and data preprocessing are mostly correct, meaningful and justified.	The chosen EDA and data preprocessing are limited and not justified.	The chosen EDA and data preprocessing are missing or incorrect.
2-5 (3%)	Statistics and Feature Engineering techniques are correct, meaningful and justified.	Statistics and Feature Engineering techniques are mostly correct and meaningful, justified.	Statistics and Feature Engineering techniques are limited and not justified.	Statistics and Feature Engineering techniques are missing or incorrect.
6-8 (8%)	All Models were identified with detailed explanations.	All models were identified and explained at a high level.	Some models were identified and explained.	Models were not properly identified and explained.
9-18 (4%)	Results of models are reported correctly and properly justified. Assumptions and Explanations are clearly and properly justified.	Results of models are reported were correct buy not properly justified. Assumptions and Explanations were clearly stated but not properly justified.	Results of models are reported correctly but incomplete and not properly justified. Assumptions and Explanations correct but are incomplete and not properly justified.	Results of models are either missing or reported incorrectly. Assumptions and Explanations are missing or incomplete.
19-20 (3%)	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is presented clearly. <u>Two (2)</u> additional steps are clearly stated and correct. Justification is complete.	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is mostly presented clearly. <u>Two (2)</u> additional steps are clearly stated and correct. Justification is mostly complete.	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is incomplete. <u>One (1)</u> additional step is clearly stated and correct. Justification is incomplete.	Conclusions are either missing or not reported correctly. Alternative steps are not identified or incorrectly stated.

Submission Format

In the DC Connect, post the following:

1. **The ran jupyter notebook file.** Any submission other than the format of a notebook file will be graded to zero.
2. A detailed word document highlighting all the work done on the project.

Academic Integrity and Late submission:

Assignments are due by the due date announced in class and posted on DC Connect. At his or her own discretion, and depending on the nature of the assignment, each professor will provide a facility for the submission of late assignments up to a maximum of 72 hours after the assignment due date. All allowed late submissions will be assessed a penalty of 25% of the total possible grade for the assignment. Assignments should be submitted on time, on a regular basis, to enable you to stay on track within the class.

Any violation of academic integrity will not be accepted and will be given a grade of zero (0) and reported. Find more information on academic integrity here <https://durhamcollege.ca/mydc/learning-resources/academic-integrity>
