**Assignment 3**

Part A: Code

One of the most useful applications of Machine Learning is to detect different diseases. In this assignment, a heart disease data set is going to be used.

The following attributes(features) are used:

1. #3 (age)  
2. #4 (sex)  
3. #9 (cp)  
4. #10 (trestbps)  
5. #12 (chol)  
6. #16 (fbs)  
7. #19 (restecg)  
8. #32 (thalach)  
9. #38 (exang)  
10. #40 (oldpeak)  
11. #41 (slope)  
12. #44 (ca)  
13. #51 (thal)

More information about the data set: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

The classification algorithm to be used will be K Nearest Neighbours (K-NN), a supervised learning method, in order to detect the presence of a heart disease.

One of the issues with these type of data sets is that some information can be useless or even can make the classifier to obtain a worst performance. For this reason, feature selection is one of the most important steps in Machine Learning.

**What to do:** Your assignment now is to apply random-restart hill climbing and random-restart Variable neighbour search as feature selection methods for the given problem. You have to run each algorithm 10 times and give the average performance. You need to present the code to the teacher. **(1 point)**

**Time to submit:** 4 April – 4 June & 27 June & 14 August

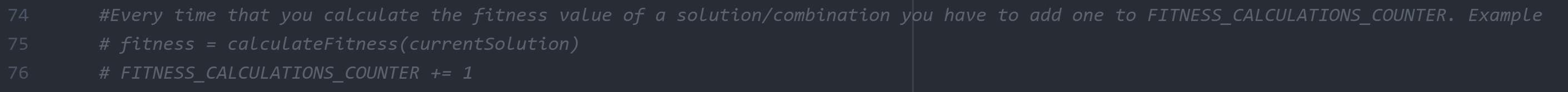
**How to use the given code:**

* Reading the dataset and K-NN is given in Assignment\_3.py
* You have to install sklearn: pip install sklearn
* Here you have to change your code:

En bild som visar text

Automatiskt genererad beskrivning

* Every time that a fitness evaluation is calculated, 1 has to be added to FITNESS\_CALCULATIONS\_COUNTER. One example can be found in the code:



* The function will stop when the FITNESS\_CALCULATION\_COUNTER is greater than MAX\_FITNESS\_CALCULATIONS (5000)

En bild som visar text

Automatiskt genererad beskrivning

* When running the algorithm, you should print something like this:

En bild som visar text

Automatiskt genererad beskrivning

Use the code given in: “Assignment\_3.py”

Before defending the code to Miguel Leon, you have to send the code by email to [Miguel.leonortiz@mdu.se](mailto:Miguel.leonortiz@mdu.se).

**I will not consider a submission without the following format:**

File names for the submission:

StudentName\_StudentLastName\_Assignment3\_Code.zip (Only by email)

Email subject for the submission:

[DVA264] StudentName StudentLastName Assignment3 Code

Part B: Report and seminar (In Swedish) 4points

I del A ovan använde du en databas över hjärtsjukdom för att göra s k ”feature selection”, dvs vilka attribut som är väsentliga för att avgöra om en person har en hjärtsjukdom eller inte. I den här deluppgiften ska du göra etisk- och könsanalys av datamängden du använde och de resultat du fick och dels skriva en kortare rapport, dels delta i diskussion vid ett obligatoriskt seminarium.

**Rapport, denna måste innehålla följande delar**

1. Diskutera bias i datamängden utifrån den litteratur du läst. Använd dig av artikeln om etik samt ”A Survey on Bias and Fairness in Machine Learning”. Vilka typer av bias kan förekomma i den här datamängden?

2. Diskutera kön i anslutning till datamängden: Är kön ett relevant attribut? Är data relevanta för både män och kvinnor? Kan de antas ge rätt resultat för båda könen? Titta t ex här på variabeln ”chest pain type”. Anser du att det går att använda databasen i ett system som ska användas för både män och kvinnor?

**Villkor för att godkänna rapporten**

Rapporten ska omfatta 400 - 600 ord (exklusive referenser) och vara skriven på svenska. Det finns ingen explicit mall för rapporten, men se till att författarnamn, student-id och titel finns med. Använd helst inte kolumner. Texten måste innehålla referenser till alla källor, inklusive kurslitteraturen. Formatera källorna enligt denna guide till IEEE-standarden <https://guides.lib.chalmers.se/ieee_guide> använd dig av fliken ”källor i referenslistan”.

För godkänt måste texten visa adekvat förståelse av ämnet och ha god struktur. Språket ska vara tydligt och korrekt. Referenser till vetenskaplig litteratur är korrekt angivna både i löptext och i referenslistan.

Rapporten ska lämnas in innan seminariets början. Se nedan om vad som händer om rapporten inte lämnats in i tid.

**Seminarium**

Förberedelse: Titta på ditt resultat i laboration 3, och svara på följande frågor: Vilka attribut (”features”) fann du var viktiga för att bedöma om personen har en hjärtsjukdom? Är ditt resultat rimligt, givet vad de olika attributen betyder? Är de samband du hittat kausala samband eller bara korrelation? Skriv gärna ner dina svar, men detta är inget du behöver lämna in.

Vid seminariet ska ni diskutera:

* Frågorna ovan om resultatet av laboration 3. samtliga ska ges tillfälle att redovisa sina svar på frågorna. Vilka olikheter finns mellan era lösningar? Finns det några ”features” som alla har fått som resultat av laborationen? Om inte – hur kan det komma sig? Diskutera särskilt också frågan om korrelation vs kausalitet
* Diskutera utifrån era rapporter vilka bias ni anser kan finnas i datamängden. Diskutera också frågorna om kön i datamängden utgående från rapporten ni skrivit.

**Villkor för godkänt på seminariet**

Seminariet har obligatorisk närvaro. Det krävs också aktivt deltagande under seminariet och att rapporten är inlämnad senast vid seminariets början. Om du inte kan delta i seminariet ska du skriva en kort rapport. Denna ska innehålla en redogörelse för det som står som förberedelser för seminariet, dvs analys av ditt resultatet från laboration 3. Denna text ska omfatta 200 – 400 ord exkl referenser. Texten skickas som pdf i email till [Christina.Bjorkman@mdu.se](mailto:Christina.Bjorkman@mdu.se) **Deadline: 4:e juni**

**Om rapporten ej lämnats in före seminariet**

Om du ej lämnat in rapporten i tid, så måste du skriva en utökad text. Rapporten ska då, utöver det som står ovan om vad den ska innehålla, också innehålla en redogörelse för diskussionen kring laboration 3 under seminariet, inkl en analys av dina egna resultat enligt vad som står under förberedelser ovan. Rapporten ska då som helhet omfatta 700 – 900 ord. **Deadline: 4:e juni**

En extra möjlighet att lämna in texterna ges 14/8. Notera: för godkänt på rapporten måste även seminariet vara godkänt! Om du varken lämnat in rapporten i tid eller deltagit i seminariet ska du kontakta Christina som beslutar om extrauppgift.

**Litteratur om datamängden och dess användning**

R Detrano, A Janosi, W Steinbrunn, M Pfisterer, J-J Schmid, S Sandhu, K H Guppy, S Lee, V Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease*”; American Journal of Cardiology*, vol 64, s. 304—310 Texten finns här på canvas. **De första sidorna i denna artikel innehåller väsentlig information om datamängden!**

S Rawat, “Heart Disease protection”, *Towards Data Science*[Online]. Aug. 2019. Tillgänglig: <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc> (hämtad 2022-05-09]

HeartDisease: Mixed data: Cleveland Heart disease data: <https://rdrr.io/cran/MixAll/man/HeartDisease.html>

Hear Disease data set: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease> Den datamängd du använt är den som heter processed.cleveland

**GRADING**

***To approve the assignment, you MUST defend the code. If the code is not defended, It does not matter the score in the report, it will be considered as fail.***

**SCORES 3-5**

* Your score < 3: U
* 3 <= your score <= 3.75: 3
* 3.75 < your score < 4.75: 4
* 4.75 <= your score: 5

**SCORES A-F**

* Your score < 3: F
* 3 <= your score < 3.4: E
* 3.4 <= your score < 3.8: D
* 3.8 <= your score < 4.2: C
* 4.2 <= your score < 4.6: B
* 4.6 <= your score: A

You are allowed to send the assignment a maximum of three times. When you approve the assignment, you are not allowed to submit the assignment again.