

Practical Data Science with Python

Assignment 2

Heart Failure Report

s3737937@student.rmit.edu.au - Huiyu Wang
s3688144@student.rmit.edu.au - Angelo Parlade
23 May, 2021

We certify that this is all our own original work. If we took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission.

Table of contents

Table of contents	2
Summary	3
Introduction	3
Methodology	4
Results	6
Discussion	8
Conclusion	8
References	9

Summary

'Heart failure', is not an independent heart disease, but a clinical syndrome of various heart diseases that have progressed to a severe stage (Medtronic, 2018). Heart failure is a serious global public health problem and is the final stage in the progression of most cardiovascular diseases (Medtronic, 2018). Most heart failure has a clear cause. To handle the heart failure issues, we create a data model from this dataset which contains medical records of 299 patients with heart failure that were collected during follow-up, with 13 clinical characteristics for each patient profile. Motivated from these facts, in this paper, we present a data model with the use of machine learning. Our contributions in this paper are in three fold. Firstly, we pre-process Heart failure clinical records Data Set from UCI webpage by retrieving and preparing the data we are going to analyse and model. Then, we illustrated each column, using appropriate descriptive statistics and graphs. Then, the relationship between 10 pairs of attributes are also discussed in the appropriate graphs. Finally, we provided two different data models by treating it as a classification task to provide deep insights to the readers about its survival rate of patients with heart failure.

Introduction

The goal of the project is ,with the use of machine learning, to create a data model in order to predict the survival rate of patients with heart failure using available clinical features or attributes of patients. Such a data model would bring the ability early prognosis of patients in order to circumvent any possible future conditions. Being able to predict future outcomes early just from a patient's current conditions, would allow you to prioritize patients at high risk rather than those at low risk which can lead to saving more lives.

By using the data set provided by the UCI Machine Learning Repository regarding heart failure clinical records, our goal is to identify the clinical features that heavily correlate with the survival rate of patients with heart failure and create a data model in order to predict future cases with a relatively high prediction accuracy rate.

Methodology

Task 1

Firstly, we start by loading the CSV data from the file, using pandas functions. And we check whether the loaded data is equivalent to the data in the source CSV file. Then, we need to clean the data. Finally, we deal with the issues or errors in the data appropriately. According to the attribute information from the official data source, age should be int, so dtype of age here we check is integer type. However, we find an error of age which is 60.667. We locate


and change them into 61. Secondly, anaemia should be boolean, so we check if the whole data 'anaemia' are either number '1' or '0'. Then, creatinine_phosphokinase(CPK)(mcg/L) should be integer as well. So, all data are positive integers. Then, diabetes should be boolean, so we check if the whole data 'diabetes' is either number '1' or '0'. Next, ejection_fraction should be percentage, so we check if the whole data is less than 100. Then, data type high_blood_pressure should be boolean, so either '1' or '0'. Moreover, platelets(kiloplatelets/mL) and serum_creatinine(mg/dL) should be float64. What's more, serum_sodium(mEq/L) should be integers. Then, sex should be binary, so we check if the whole data 'sex' are either number '1' or '0'. And, smoking should be boolean, so we check if the whole data 'smoking' is either number '1' or '0'. Finally, time should be integer. All data are positive integers.

Task 2

In the first part, we explore each column using appropriate statistics and graphs.

Firstly, we explore each column, using descriptive statistics and graphs. When we explore the column 'age', drawing the histogram for 'age'. Simply because it is clear to explore the most age range frequency and the least. Also, we explore the column 'age', drawing the density plot for 'age'. Simply because it is clear to can also figure out whether there are distributions peaks or valleys of 'age'. Secondly, we explore the column 'anaemia', draw the pie chart for 'anaemia'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart. Thirdly, explore column 'creatinine_phosphokinase(CPK)', drawing the density plot for 'creatinine_phosphokinase'. Simply because it is clear to can also figure out whether there are distributions peaks or valleys of 'CPK'. Fourthly, we explore the column 'diabetes', drawing the pie chart for 'diabetes'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart. Fifthly, we explore the column 'ejection_fraction', this data is percentage. So we need to know distributions peaks or valleys. Sixthly, we explore the column 'high_blood_pressure', drawing the pie chart for 'high_blood_pressure'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart. Seventhly, we explore the column 'platelets'. We need to know distribution peaks or valleys, drawing the density plot for 'platelets'. Eighthly, we explore the column 'serum_creatinine', drawing the histogram for 'serum_creatinine'. Simply because it is clear to explore the most serum_creatinine range frequency and the least. Ninthly, we draw the histogram for 'serum_sodium'. Simply because it is clear to explore the most serum_sodium range frequency and the least. Tenthly, we explore column 'sex', drawing the pie chart for 'sex'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart. Eleventhly, we explore the column 'smoking', drawing the pie chart for 'smoking'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart. Then, we draw the histogram for 'time'. Simply because it is clear to explore the most time frequency and the least. Finally, we explore the column 'DEATH_EVENT', drawing the pie chart for 'DEATH_EVENT'. Simply because it is boolean, it is easy to show the illustrated numerical proportion by pie chart.

In the second part, we explore the relationship between 10 pairs of attributes. Firstly, we explore the relationship between 'age' and 'anaemia'. We plot a boxplot of age by anaemia. Secondly, we explore the relationship between 'age' and 'diabetes'. We plot a boxplot of age by diabetes. Thirdly, we explore the relationship between 'age' and 'high_blood_pressure'. We plot a boxplot of age by 'high_blood_pressure'. Fourthly, we explore the relationship between 'age' and 'DEATH_EVENT'. We plot a boxplot of age by 'DEATH_EVENT'. Fifthly,



we explore the relationship between 'age' and 'creatinine_phosphokinase'. We plot a boxplot of age by 'creatinine_phosphokinase'. Sixthly, we explore the relationship between 'age' and 'ejection_fraction'. We plot a boxplot of age by 'ejection_fraction'. Then, we explore the relationship between 'age' and 'platelets'. We plot a boxplot of 'platelets' by age. Moreover, we explore the relationship between 'ejection_fraction' and 'DEATH_EVENT'. We plot a boxplot of 'ejection_fraction' by 'DEATH_EVENT'. What's more, we explore the relationship between 'serum_creatinine' and 'DEATH_EVENT'. We plot a boxplot of 'serum_creatinine' by 'DEATH_EVENT'.

Finally, we explore the relationship between 'serum_sodium' and 'DEATH_EVENT'. We plot a boxplot of 'serum_sodium' by 'DEATH_EVENT'.

Task 3


We chose to model the data as a Classification task. One of the data models uses the K-Nearest Neighbors Algorithm while the other uses the Decision Tree Algorithm. The main reason we chose to go with the classification approach was because with our chosen data set, Heart Failure Reports, made use of the predefined class/target, death event. This attribute according to the source of the data set, UCI Machine Learning Repository, corresponds to if the patient died during the follow-up period (boolean).

KNeighborsClassifier: Parameter Tuning

After multiple attempts of changing the available parameters of the classifier, such as weights, algorithm, power, metrics, etc., no noticeable increase in the model's accuracy was found except when the value of "k" was changed. This value represents the number of neighbors to be used for kneighbors queries. A for loop was used to test all odd numbers from 1 -100 as the value of "k" to see which gave the highest accuracy rating with the lowest value possible. Only odd numbers were used because there are only 2 or an even number of possible target values which was either 0 or 1 in the death event attribute. This is to avoid cases 99% of the time where the number of neighbors are equal between the 2 possible values.

KNeighborsClassifier: Feature Filtering

The Hill Climbing approach was used for feature filtering the data frame in order to improve the data model's accuracy. According to Wikipedia 2021, this optimization technique makes use of an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by making an incremental change to the solution. The arrangement of the features of the data frame is first randomly shuffled then only the first feature in the newly arranged feature list is used to test the model's accuracy. Once tested, the next feature is added to the data model and its accuracy is compared to its previous value. If the accuracy did not improve, then the last added feature is removed from the data model and the next feature is used



for the next comparison. This is done until all features are tested and will result to the best accuracy possible using the new arrangement of features.

After multiple testing attempts with using different feature arrangements, it was found that it is possible to get a higher accuracy or the same accuracy but with more features involved depending on the arrangement. So ideally, all possible permutations of the features should be tested in order to identify the best accuracy available but due to the lack of time and computational resources, we decided to make use of only 1250 permutations, which is far from the possible 490M possible permutations because of having 12 features available. It was also decided that, even though a similar accuracy rating was possible with as little as one feature involved, we prioritized data models that made use of the most number of features while maintaining the highest accuracy rating possible. While having more data does not necessarily always help, it does avoid having a model that is too simple and may be seen as highly biased. On the other hand, having too many features can lead to a model that is too complicated and lead to a high variance situation which in turn can lead to model overfitting. But because having more features was still possible while maintaining the best accuracy rating, we decided to keep the model the highest number of features.

DecisionTreeClassifier: Parameter Tuning

As we did with the KNeighborsClassifier, we tested modifying the parameters of the Decision Tree Classifier in order to get a more accurate rating. No massive amounts of testing was needed because only 2 parameters showed positive effects to the accuracy which were the criterion and min_samples_split parameters.

DecisionTreeClassifier: Feature Filtering

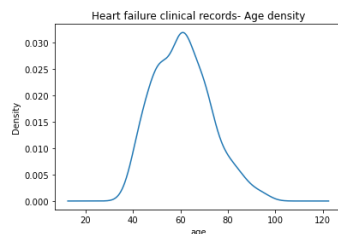
The Hill Climbing approach was used again for the feature filtering of the Decision Tree Classifier. This also meant that multiple permutations were also tested in order to find the best accuracy. The main differences between the feature filtering of this classifier compared to the last one is that because every time the data model is trained had returned different values with +-1 differences between each other, an average of 10 training attempts was kept instead for comparison with other accuracy ratings. Because this significantly added to the processing time for each loop, the range was changed from 1250 to 500 in order to save time. Another difference is that before we prioritized data models that made use of the most amount of features but still maintained the best accuracy rating recorded, this time we are prioritizing data models that have the least number of features but still maintained the highest accuracy rating recorded. This is because we wanted to avoid a massive visualization of the decision tree.

Results

Task 2

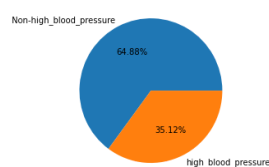
In the first part of task2, the results are discussed below.

Firstly, we can observe from the density plot of age that the middle age of heart failure clinical is about 60.

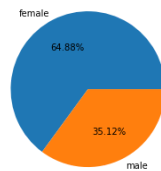


Secondly, we can observe from the pie chart of anaemia that more percentage of patients, 58.86%, are not anaemia. Thirdly, we can observe from the histogram chart of 'creatinine_phosphokinase' that the most frequent frequency of 'creatinine_phosphokinase' is lower than 1000 mcg/L. Fourthly, we can observe from the pie chart of diabetes that more percentage of patients, 58.19%, are not diabetes. Fifthly, we can observe from the density plot of 'ejection_fraction' that the most percentage of 'ejection_fraction' is near 40 percent. Sixthly, we can observe from the pie chart of 'high_blood_pressure' that more percentage of patients, 64.88%, are not high_blood_pressure.

(Figure_6)- Percentage of high_blood_pressure

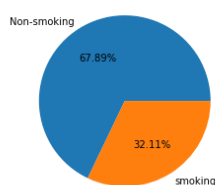


(Figure_10)- Percentage of sex

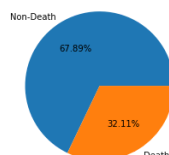


Seventhly, we can observe from the density plot of 'platelets' that the most percentage of 'platelets' is near 25000 kiloplatelets/mL. Then, we can observe from the histogram chart of 'serum_creatinine' that the most frequent frequency of 'serum_creatinine' is near 1 mg/dL. Moreover, we can observe from the histogram chart of 'serum_sodium' that the most percentage of 'serum_sodium' is about 135 mEq/L. Tenthly, we can observe from the pie chart of sex that 64.88% are female.

(Figure_11)- Percentage of smoking



(Figure_13)- Percentage of DEATH_EVENT

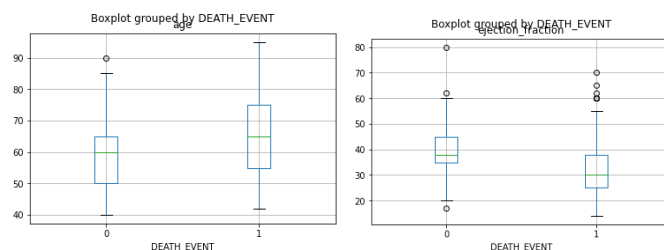


Then, we can observe from the pie chart of smoking that most percentage of patients are not smoking.

Moreover, we can observe from the histogram chart of time that the most percentage of time is about 100 days. Finally, we can observe from the pie chart of 'DEATH_EVENT' that the most percentage of DEATH_EVENT is non-death.

In the second part of task2, the results are discussed below.

Firstly, we explore the relationship between 'age' and 'anaemia'. The hypothesis is, lack of relationships. Secondly, we explore the relationship between 'age' and 'diabetes'. The hypothesis is ,the younger the patient's age, the higher chance they will have diabetes. Thirdly, we explore the relationship between 'age' and 'high_blood_pressure'. The hypothesis is ,the older the patient's age, the higher chance they will have high blood pressure. Fourthly, we explore the relationship between 'age' and 'DEATH_EVENT'. The hypothesis is ,the older the patient's age, the higher chance they will have heart failure. Fifthly, we explore the relationship between 'age' and 'creatinine_phosphokinase'. The hypothesis is lack of relationships. Sixthly,we explore the relationship between 'age' and 'ejection_fraction'. The hypothesis is, the younger the patient's age, the higher percentage of ejection_fraction.



Then,we explore the relationship between 'age' and 'platelets'. The hypothesis is, lack of relationships.

Moreover, we explore the relationship between 'ejection_fraction' and 'DEATH_EVENT'. The hypothesis is, the lower the percentage of ejection_fraction, the higher chance they will die due to heart failure.

What's more, we explore the relationship between 'serum_creatinine' and 'DEATH_EVENT'. The hypothesis is, the higher serum_creatinine in the blood, the higher chance they will die due to heart failure. Finally, we explore the relationship between 'serum_sodium' and 'DEATH_EVENT'. The hypothesis is, the lower serum_sodium in the blood, the higher chance they will die due to heart failure.

Task 3

KNeighborsClassifier: Final First Classification Model

After parameter tuning the KNeighborsClassifier, the best “k” value we found was 33. This resulted with an accuracy of 73%. This value was then used for all future KNeighborsClassifier training and testing. The best combination of features that resulted with an even higher accuracy of 88% is the use of the following features: 'smoking', 'serum_creatinine', 'serum_sodium', 'ejection_fraction', 'sex', 'age', 'time'.

This is the classification report of the best KNeighborsClassifier Model we came up with:

	precision	recall	f1-score	support
0	0.87	0.98	0.92	55
1	0.92	0.60	0.73	20
accuracy			0.88	75
macro avg	0.90	0.79	0.83	75
weighted avg	0.88	0.88	0.87	75

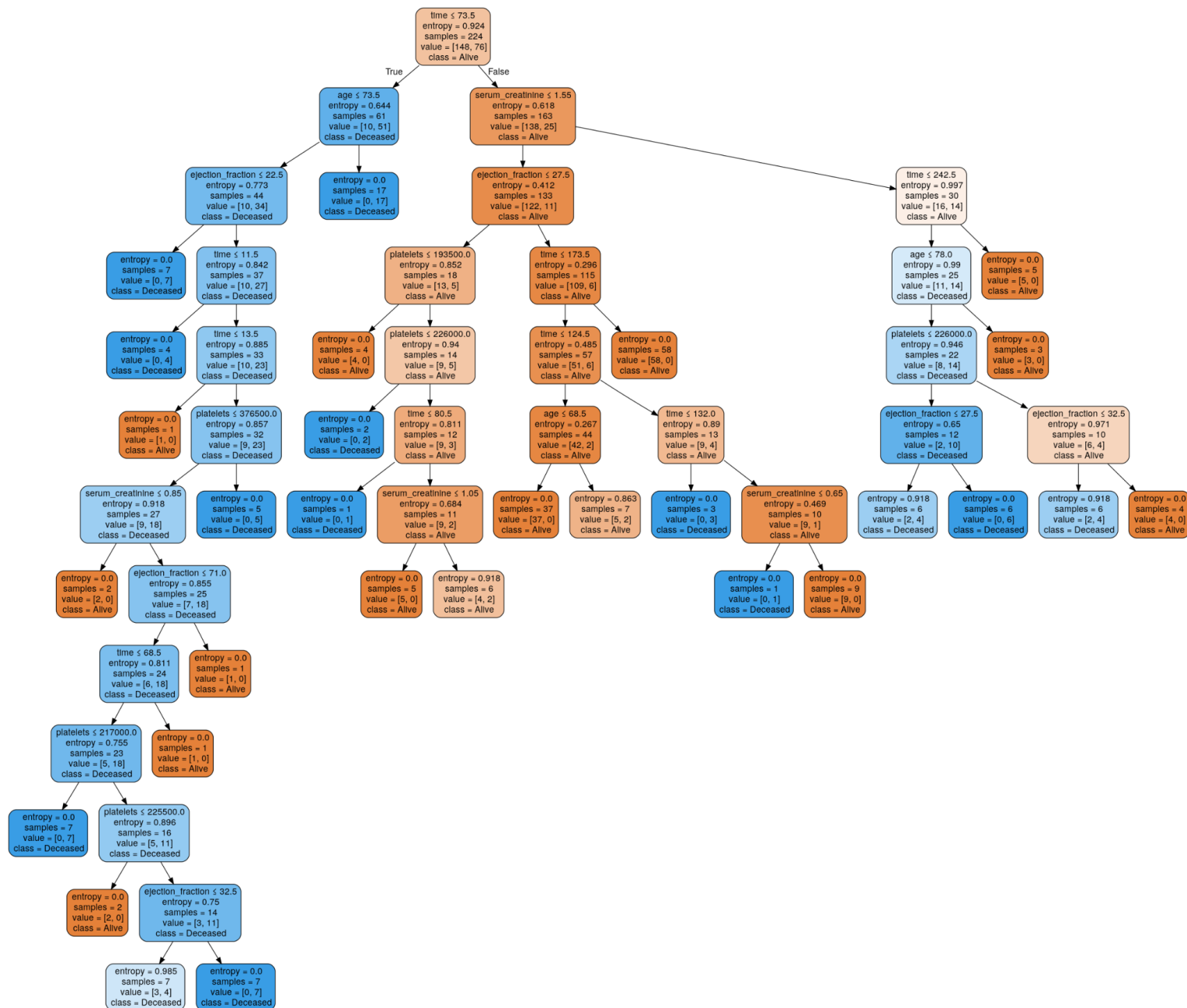
DecisionTreeClassifier: Final Second Classification Model

After parameter tuning the DecisionTreeClassifier, the best parameters we found that gave us the highest accuracy rating recorded is a criterion parameter set to 'entropy' and the 'min_samples_split' set to 8. This resulted with an accuracy of 83%. These parameters were then used for all future DecisionTreeClassifier training and testing. The best combination of features that resulted with an even higher accuracy of 91% is the use of the following features: 'platelets', 'time', 'ejection_fraction', 'serum_creatinine', 'age'.

This is the classification report of the best KNeighborsClassifier Model we came up with:

	precision	recall	f1-score	support
0	0.93	0.95	0.94	55
1	0.84	0.80	0.82	20
accuracy			0.91	75
macro avg	0.89	0.87	0.88	75
weighted avg	0.91	0.91	0.91	75


Decision Tree Visualization



Discussion

Task 2

Main age in heart failure clinical patients is about 60. However, the older the patient's age, the higher chance they will have to die due to heart failure. Moreover, the higher percentage



of ejection fraction will also cause death. Serum_sodium and serum_creatinine in the blood may cause death as well.

Task 3

The KNeighbors Classifier Model was able to identify the correlation of between heart failure deaths and the following attributes; 'smoking', 'serum_creatinine', 'serum_sodium', 'ejection_fraction', 'sex', 'age', 'time'; with an accuracy rating of 88%. While the Decision Tree Classifier was able to identify the correlation between heart failure deaths and the following attributes; 'platelets', 'time', 'ejection_fraction', 'serum_creatinine', 'age'; with an accuracy rating of 91%. The common attributes found in both models are 'time', 'age', and 'serum_creatinine'. This may represent that these attributes contribute the most to their relation with heart failure deaths. While the other attributes are used but not found in both models, does not mean that they are not relevant to heart failure but can instead further correlate with other data found in their respective models. Another observation upon further testing, the accuracy rating fluctuates between 91% and 89%. Both percentages are still higher than the accuracy rate of the KNeighbors classification. This together with the fact that Decision Tree classifications support automatic feature interaction, whereas KNeighbors classifications can't and that they are faster due to KNeighbors classifications' expensive real time execution has shown to us which of the two models should be used.

Conclusion

Task 2

The graphs we recommend to use are boxplot, histogram and pie chart. These graphs can represent the relationship between death and the other 12 attributes.

Task 3

The model we recommend to use is the Decision Tree classification due to its higher accuracy rating and less expensive real time execution. This model represents the correlation of heart failure deaths to 'platelets', 'time', 'ejection_fraction', 'serum_creatinine', and lastly 'age'.

References

- Gorini, M. (2021). *Classification Vs. Clustering - A Practical Explanation*. Bismart.
<https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation#:~:text=Although%20both%20techniques%20have%20certain,which%20differentiate%20them%20from%20other>
- Navlani, A. (2018, August 3). *KNN Classification using Scikit-learn*. Datacamp.
<https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>
- sklearn.neighbors.KNeighborsClassifier* — *scikit-learn 0.24.2 documentation*. (2021). Scikit Learn.
<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
- Varghese, D. (2019, May 10). *Comparative Study on Classic Machine learning Algorithms*. Medium.
<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=Decision%20tree%20vs%20KNN%20%3A,KNN's%20expensive%20real%20time%20execution>
- Wikipedia contributors. (2021, May 3). *Hill climbing*. Wikipedia.
https://en.wikipedia.org/wiki/Hill_climbing
- Medtronic. (2018). *What is Heart Failure?*.
<https://www.medtronic.com/au-en/patients/conditions/heart-failure.html>