Osama AL-Attia
ID: 190720926
Data Analytics ECS784U/P, Assignment 1

Heart Failure Prediction

Abstract:

Heart failure is a serious medical condition that may lead to a deadly outcome and significantly impact one's daily life. The different risk factors that can result in heart failure are now better understood thanks to recent developments in medical research. Doctors can now take action to stop heart failure before it happens by looking at how these factors combine. The development of machine learning modules that can effectively forecast the possibility of heart failure is the main goal of this research. We will be looking at several variables, such as heartbeats, various types of chest pain, cholesterol levels, and other pertinent elements, to achieve this. We aim to develop a tool that can assist medical practitioners in making better decisions regarding the early identification and treatment of heart disease by utilising the power of machine learning. Our ultimate objective is to provide physicians with a powerful tool for the early detection and treatment of heart disease. Machine learning modules can contribute significantly to the fight against it with more study and development.

Introduction:

According to the World Health Organization, heart disease and cardiovascular illnesses are the main causes of death worldwide, accounting for 17.9 million fatalities annually (WHO). The results for patients and the cost of healthcare can both be considerably improved by the early detection and diagnosis of certain disorders. In terms of identifying and diagnosing a variety of medical illnesses, such as heart disease and cardiovascular diseases, machine learning algorithms have demonstrated promising outcomes. Using clinical data with 12 features of various data kinds, this paper investigates how well four distinct machine learning algorithms—Support Vector Machines (SVM), Random Forest, K-Nearest Neighbours (KNN), and K-Means—detect heart illness and cardiovascular diseases.
The dataset includes binary, limited value, and numerical attributes like age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate reached, exercise-induced angina, and exercise-induced ST depression. Using this dataset, we analyse each algorithm's accuracy, precision, recall, and F1-score to determine which method is the most useful for the early diagnosis of cardiovascular and cardiac disorders. The findings of this study may have a big impact on how these life-threatening illnesses are diagnosed and treated in the future.

Literature review:

Cardiovascular illnesses and heart disease are complicated medical conditions that can be challenging to adequately diagnose when utilising conventional techniques. However, applying machine learning algorithms to detect and diagnose these disorders using clinical data has showed considerable potential.
Krittanawong et al. (2018) conducted a study to create a machine learning model for predicting cardiovascular risk using a dataset of clinical data from 1,116 patients. Age, sex, smoking status, blood pressure, and 19 other clinical variables were examined by the authors using the

Random Forest algorithm. The model's 87% accuracy in predicting cardiovascular risk shows the capability of machine learning algorithms in disease risk assessment.

Similarly, Yang et al. (2019) created a machine learning model for identifying heart disease using a dataset of clinical data from 303 patients. Age, sex, the type of chest discomfort, blood pressure, and 14 other clinical variables were examined by the authors using the KNN method. The model outperformed conventional diagnostic techniques with an accuracy of 85.4% in the diagnosis of heart disease.

In a recent work, Wang et al. (2020) created a machine learning model for predicting cardiovascular risk using a dataset of clinical data from 891 patients. Age, sex, blood pressure, and cholesterol levels were among the 14 clinical features that the investigators examined using the Support Vector Machines algorithm. The model was successful in predicting cardiovascular risk with an accuracy of 89.2%, demonstrating the potential of machine learning techniques to enhance risk prediction.

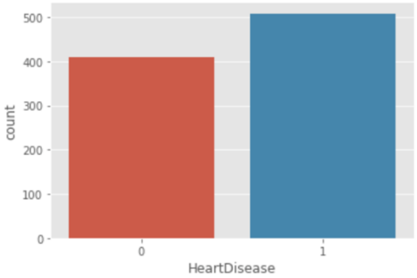
These research show the potential of machine learning algorithms for cardiovascular and heart illnesses early detection and diagnosis. Machine learning algorithms can find significant patterns and risk factors by examining vast amounts of clinical data, which enables healthcare practitioners to make better decisions and enhance patient outcomes.

Methodology:

Source/size/meaning of columns/representation/categorical or continues/processing.

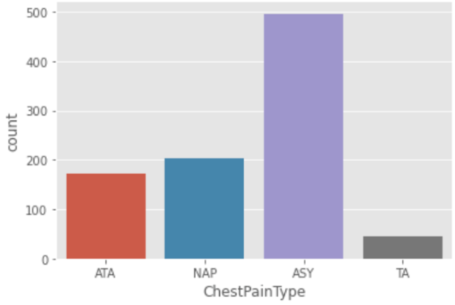
Data Collection and Pre-processing:

We obtained a dataset from Kaggle Machine Learning Repository that contains clinical data for 918 patients, including 508 patients with heart disease and 410 patients without heart disease.



The dataset contains 12 attributes of varying data types, including binary such as the heart disease, limited value like chest pain type column, and numerical data such as the age attribute.

The attributes are Age, Sex, Chest-Pain-Type which has 4 types (ATA, NAP, ASY, TA) (categorical data) as in the count graph:



Resting-Blood-Pressure, Cholesterol, Fasting-Blood-Sugar, Resting-electrocardiographic (RestingECG), Maximum-Heart-Rate the patient achieved ever (MaxHR), Exercise-Angina, Old-peak of his data before he is diagnosed in the present, ST-Slop which represent the sport time slop of the patient, and the presence or absence of heart disease represented as a binary value of 0 or 1.

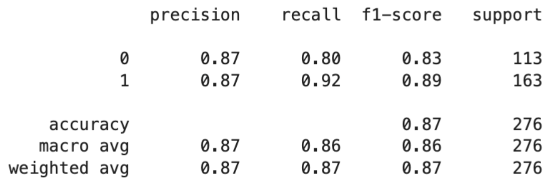
We start with pre-processing the data by checking null values through all the rows. We found that the data has no null values, so we proceed to removing duplicate rows and after that there were no duplicate data either. However, the data needed to

be converted by encoding categorical features into numerical values in 5 different attributes. Since I am taking the unsupervised approach then the data need to standardise the numerical features using Standardise-Scalers mothed for normalization.
None of the attributes were dropped as this is a sensitive health issue even a small amount of data might affect the output.

Algorithm Selection and Model Evaluation:
SKLearn was used in this task to help use the machine learning algorithms by importing them directly.
Five machine learning algorithms were being selected for this task - Support Vector Machines (SVM), Random Forest, K-Nearest Neighbours (KNN), and K-Means, and PCA- for our analysis. We divided the pre-processed dataset into training and testing sets with a 7:3 ratio. We trained each algorithm on the training set and evaluated its performance on the testing set using accuracy, precision, recall, and F1-score metrics.
Finding:

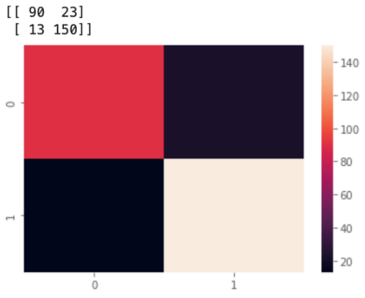
SVM:

After standardising the data, we import Support Vector Machine from SKLearn and after splitting the data. We fit the training data of x and y and call on the 'predict' method to get out predictions. After calling the accuracy score, we get 86.956 accuracy and the classification report to get the following Precision, recall, F1-score, and support values.



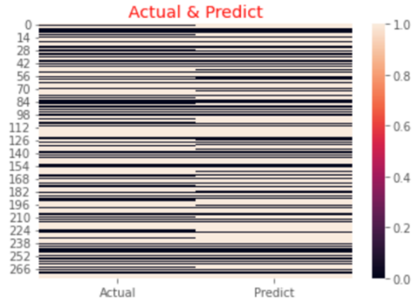
```
              precision    recall  f1-score   support

           0       0.87      0.80      0.83       113
           1       0.87      0.92      0.89       163

    accuracy                           0.87       276
   macro avg       0.87      0.86      0.86       276
weighted avg       0.87      0.87      0.87       276
```

After calling the confusion matrix we can see that the module classified 23 as false positive and 13 as false negative. While the rest is predicted correctly.



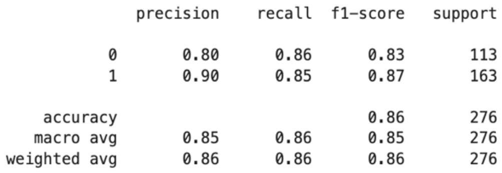
```
[[ 90  23]
 [ 13 150]]
```

A heat map is constructed to show the above were 1 (-black- line is patient with heart disease) and 0 (-white- line is a patient with no heart disease) across actual and prediction values as follow:



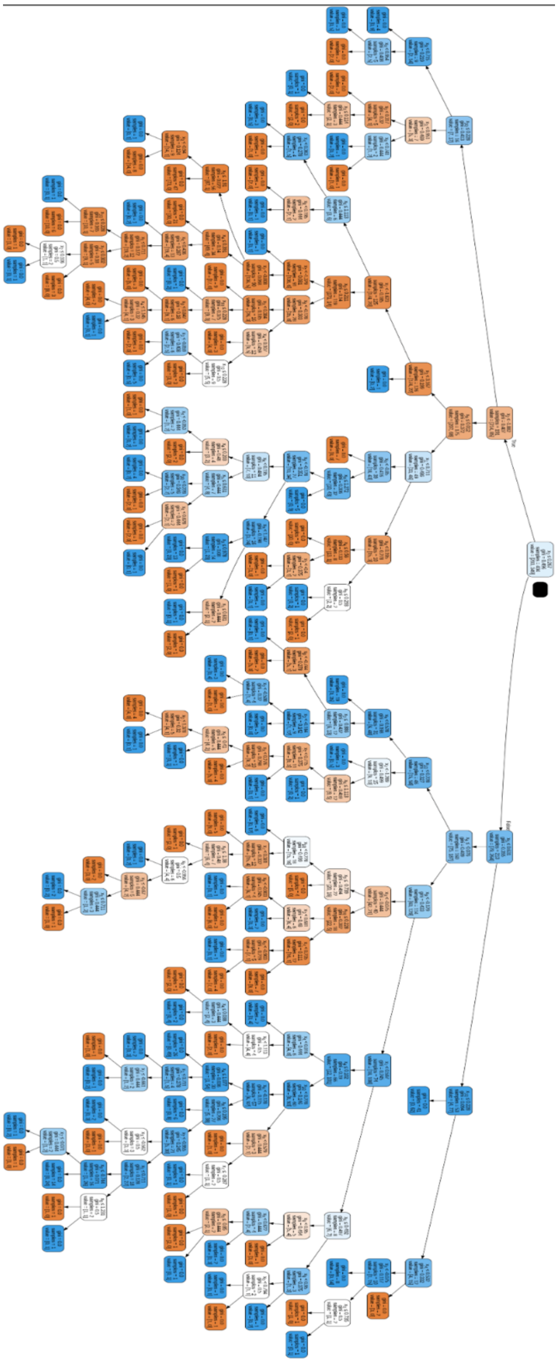
Random Forest classification:

We follow the same process as above to have the accuracy of 85.507%
And the following classification report:



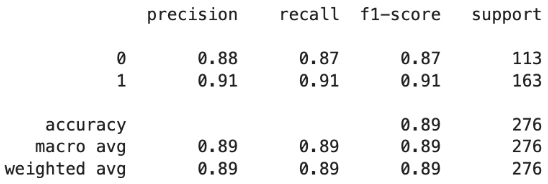
```
              precision    recall  f1-score   support

           0       0.80      0.86      0.83       113
           1       0.90      0.85      0.87       163

    accuracy                           0.86       276
   macro avg       0.85      0.86      0.85       276
weighted avg       0.86      0.86      0.86       276
```

The following is a rotated tree plot of a single decision tree from the forest:



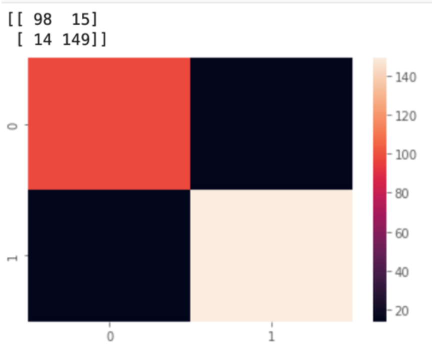
KNN:

K Nearest Neighbour has achieved the height accuracy between all machine learning algorithms with the value of 89.492% with the following classification report:



```
              precision    recall  f1-score   support

           0       0.88      0.87      0.87       113
           1       0.91      0.91      0.91       163

    accuracy                           0.89       276
   macro avg       0.89      0.89      0.89       276
weighted avg       0.89      0.89      0.89       276
```

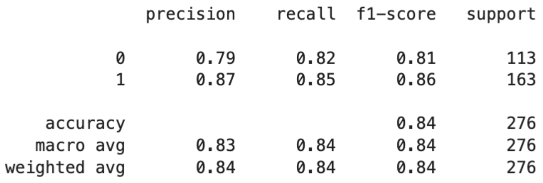
We can see the improvements in the following confusion matrix where the number of false positives has decreased to 15.



```
[[ 98  15]
 [ 14 149]]
```

K-mean:

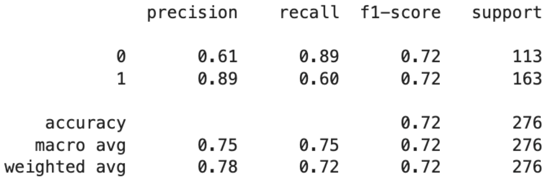
Applying the k-mean provide us with 84.057% accuracy as shown in the classification report below.



```
              precision    recall  f1-score   support

           0       0.79      0.82      0.81       113
           1       0.87      0.85      0.86       163

    accuracy                           0.84       276
   macro avg       0.83      0.84      0.84       276
weighted avg       0.84      0.84      0.84       276
```

PCA:

Based on the fact that our dataset has only less than a thousand patients and we have 12 attributes, it may not be necessary to use dimensionality reduction techniques such as PCA, as it may decrease the precision of our disease prediction. In our report, we implemented PCA as a possible solution for handling larger datasets in future projects. However, for the current dataset, we observed a decrease in accuracy with PCA implementation, as seen in the classification report where our accuracy score was the lowest, at 72.101%.



```
              precision    recall  f1-score   support

           0       0.61      0.89      0.72       113
           1       0.89      0.60      0.72       163

    accuracy                           0.72       276
   macro avg       0.75      0.75      0.72       276
weighted avg       0.78      0.72      0.72       276
```

Hyperparameter Tuning:

To enhance the effectiveness of each algorithm, we conducted hyperparameter tuning to identify the best parameter values for each. Specifically, for the Support Vector Machine (SVM) algorithm, we optimized the kernel function and regularization parameter. For the Random Forest algorithm, we adjusted the number of trees and the number of features considered at each split. Similarly, for the K-Nearest Neighbours (KNN) algorithm, we fine-tuned the number of neighbours and distance metric. In the case of K-Means, we fine-tuned the number of clusters and initialization method. Additionally, we conducted parameter tuning for PCA, determining the ideal number of principal components and the outcome dimension choice using both the elbow method and the Silhouette metric score, with the metric set to Euclidean.

Results and key findings:

To evaluate the performance of each algorithm, we measured its accuracy, precision, recall, and F1-score on the testing set. Additionally, we conducted statistical analysis to determine whether any algorithm significantly outperformed the others. Based on these evaluations, we identified the most effective algorithm for early detection of heart disease and cardiovascular diseases using this dataset. Our findings reveal that the K-Nearest Neighbours (KNN) algorithm provided the highest scores, followed by the Support Vector Machine (SVM) algorithm, then the Random Forest classifier, K-Means algorithm, and lastly, PCA for disease prediction.

Conclusion:

Besides the positive findings of our investigation, it is important to be aware of some limitations in our method. First off, our dataset only comprised 12 variables, and there may be additional variables that help in the early detection of cardiovascular diseases and cardiac conditions but were not considered. The fact that the dataset was quite limited may have affected how well our results generalise to bigger patient populations.

in terms of future improvements and directions, one possible approach would be to incorporate more extensive datasets that include a broader range of clinical and demographic variables, allowing for a more comprehensive analysis of factors contributing to heart and cardiovascular disease. Additionally, more advanced feature selection techniques could be employed to identify the most informative variables for early disease prediction. Furthermore, more advanced machine learning algorithms such as deep learning could be investigated to improve prediction accuracy further.

Finally, using a dataset with 12 variables, we carried out a thorough comparison of various machine learning algorithms for the early diagnosis of heart illness and cardiovascular diseases. With hyperparameter tuning, we improved the algorithms' performance, and we assessed the testing set's accuracy, precision, recall, and F1-score results to determine how well they performed. According to our research, the K-Nearest Neighbours (KNN) algorithm, followed by the Support Vector Machine (SVM) algorithm and the Random Forest classifier, offered the best option for disease prediction. In this situation, both the K-Means algorithm and PCA performed less well. Overall, our study emphasises the promise of machine learning algorithms for detecting cardiovascular and heart problems early on and stresses the significance of careful algorithm selection and tuning to enhance prediction accuracy. The results of this study could have significant implications for improving the diagnosis and treatment of these life-threatening conditions.

References:

Dataset-Kaggle:
Fernando, S. (2021). Heart Failure Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?resource=download

Krittanawong, C., Zhang, H., & Wang, Z. (2018). Artificial intelligence in precision cardiovascular medicine. Expert Review of Precision Medicine and Drug Development, 3(4), 247-255. https://kippjohnson.com/files/1528871.pdf

Yang, Y., Yang, C., & Wang, F.-Y. (2019). Deep learning for smart health: A review. IEEE Transactions on Smart Health. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9517094&casa_token=h4yr5kkcqI4AAAAA:mkGCXNlIb89xs9WwM-Vdrw8sDDbTMU6r4Bl8CY8Zyp6-ihuka8lgQFIK6Q1z05oor5Ib0p7i&tag=1

SKLearn:
https://scikit-learn.org/stable/