**Department of Computer Science**

**Forman Christian College University**

**COMP360: Introduction to AI**

**Spring 2022**

Title:

Heart Disease Predictor

**Shaheer Ur Rehman 231451798**

**Muhammad Arham 231453689**

Logo, company name

Description automatically generated

Abstract:

We developed some intelligent models which tell us whether someone has heart disease or not based on some data like their age, cholesterol level, etc. The idea was that some attributes of a person like resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate achieved, etc. dictate whether that given person has heart disease. After finding the data online, we developed Naïve Bayes and K-Nearest Neighbor programs to run on that data set.

Introduction:

Heart diseases are one of the main reasons for death among the human species. Timely diagnosis of such diseases can help us to save many lives. A lot of lives are lost just because heart disease was not diagnosed on time. To get a diagnosis, patients must go through some very expensive procedures which also take a lot of time. What if there existed a software that could guess if you have heart disease or not so that if you get a positive result, you are surer that you need to get yourself tested.

Literature Review:

# 1. Research Paper 1

## Title:

Artificial Intelligence, Machine Learning, and Cardiovascular Disease

## Reference:

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial Intelligence, Machine Learning, and Cardiovascular Disease. Clinical Medicine Insights: Cardiology. <https://doi.org/10.1177/1179546820927404>

According to research done in 2020, AI is implemented in lots of medical fields to detect patterns that could be indicative of heart disease, lung disease, and cancer. Given the first computer that was used for testing pulmonary function patients with lung disease. Most of the functions that these computers used were Boolean logic, not Artificial intelligence. Now, we are at a point where we look at big data sets of already sick and healthy patients, to make a model to learn to check diseases. They look at various medical tests such as CT scans MRI scans, and X-rays and detect diseases. The paper discusses how the brain and deep learning model are similar in ways that they are composed of nodes that accept input and give output. Many algorithms can use for deep learning. Such as linear regression, logistic regression, and decision trees. The cardiovascular disease detection system studied in this paper looks at the EKG of the patient. The model was first trained with data around 45k patients and 55k patients of unique data sets. The model was trained using CNN deep learning method. In detail, the model was trained with 44959 patients with 12 lead EKG and echocardiogram data set as the ventricular dysfunction. The model was trained with independent data set of 55 870 patients. This is a vast data set that is more than enough to train a model but in medicine, more accuracy is always welcome. The CNN model in this study showed an accuracy of 85.7%.

# 2. Research Paper 2

## Title:

Supporting Real World Decision Making in Coronary Diseases Using Machine Learning

## Reference:

Kokol, P., Jurman, J., Bogovič, T., Završnik, T., Završnik, J., & Blažun Vošner, H. (2021). Supporting Real World Decision Making in Coronary Diseases Using Machine Learning. INQUIRY: The Journal of Health Care Organization, Provision, and Financing. <https://doi.org/10.1177/0046958021997338>

According to this research paper from 2020/2021, there have been a plethora of studies reported with high accuracy rates of machine learning algorithms (MLAs) in medical applications, and most of these studies used cleaned medical databases without the presence of the “real world noise.” But in this study, the aim was to perform machine learning on the routinely collected Anonymous Cardiovascular Database (ACD), extracted directly from a hospital information system of the University Medical Centre Maribor. Another thing regarding methodology and accuracy, that is used in this paper, is that the study was performed in 2 phases: In the first phase, different MLAs were trained on a comparable University of California Irvine (UCI) Heart Disease Dataset. This phase aimed to define the “standard” ACU values and second, to reduce the set of all MLAs to the most appropriate candidates to be used on ACD, during the second phase. Using this technique, some of the MLAs were selected. These MLAs achieved around 0.96 accuracies on the ACD. The accuracy of the ACD (real world and not “noise-free” clinical data) reached the highest levels using decision trees and neural networks while Linear regression and AdaBoost performed best in the UCI database. There have been some limitations as well. The datasets used in this study varied in the presentation of data. The UCI dataset was significantly smaller and only patients with cardio diseases were included. Therefore, there were a small number of samples and a limited set of attributes. Contrary, there were no healthy patients in the ACP database, consequently, the generation of healthy individuals might have influenced the classification results. The accuracy of machine learning differs significantly between the 2 databases, indicating that decision tree-based algorithms and neural networks might be better at coping with real-world not “noise-free” clinical data.

# 3. Research Paper 3

## Title:

Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype

## Reference:

Russak, A. J., Chaudhry, F., De Freitas, J. K., Baron, G., Chaudhry, F. F., Bienstock, S., Paranjpe, I., Vaid, A., Ali, M., Zhao, S., Somani, S., Richter, F., Bawa, T., Levy, P. D., Miotto, R., Nadkarni, G. N., Johnson, K. W., & Glicksberg, B. S. (2020). Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype. Journal of Cardiovascular Pharmacology and Therapeutics, 379–390. <https://doi.org/10.1177/1074248420928651>

According to this study conducted on May 3, 2020, given how much artificial intelligence has been engrossed in the medicine market, it is predicted that by 2026 it will be a 150-billion-dollar market. In the last years of research Machine learning in medicine has come a long way. We have been accurately able to characterize vertebral compression fractures, Tuberculosis long lesions, and lung nodules on imaging studies. The research also has been used to detect skin cancer and breast cancer. This is just one of the few fields in machine learning that has helped to identify or detect disease. We still have a long way to go in cardiovascular disease or heart disease detection using machine learning. This study also explains the importance of unbiased data and how bias especially in medicine can absurdly make a model useless. That’s why for their testing they have chosen a wide range of patients for their data set. Using imaging software such as EKG, MRI, and x-rays among many patients who have been diagnosed with heart failure or abnormalities or those undergoing procedures, added perfectly healthy lifestyles to more sedentary lifestyles. This way of collecting data will reduce bias in machine learning algorithms. Also, this study mentions the use of cloud computing to mix data sets and achieve much better accuracy in the model. It is noted that no matter how accurate the model becomes it shouldn’t be used alone to diagnose just to classify among the masses.

Methodology:

We used naïve Bayes classification which is a probabilistic classifier, and it classifies based on the Bayes theorem with strong (naïve) independence assumptions between features. In our data, there was some continuous data that we had to delete, to make naïve Bayes work. Naïve Bayes is a supervised learning algorithm

We also used K-Nearest Neighbor (KNN) classification which classifies based upon the distance from K. It is also a supervised learning algorithm in which the input consists of the k closest training examples in a data set.

Furthermore, we used K-means clustering to represent the data. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

Experiment & Results:

# Naïve Bayes classifier:

For the naïve Bayes classifier, We are using gausisianNB in python library SK learn to train and test the data.

Naïve Bayes works best with noncontinuous data ( for example in the values of YES and No)

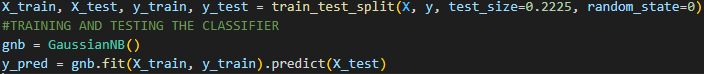


We have to drop chol, trestbps, fbs, thalach and old peak in the data for the classifier to work best.

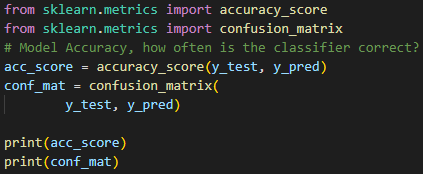
Here, We are dropping the data.



Now, we have to divide the data for “training” the classifier and “testing” the classifier. Here we train and test the data(around 22% of rows from the original rows are used for testing and are not included in the training of the data from a total of 1300 rows).



Now we need to get the accuracy and confusion matric of the classifier to know how well it is working.



We are using the sklearn metrics library to evaluate the classifier performance.



Here, we can evaluate the accuracy of the classifier is around 84%.

From the given values we can construct the confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| *Predicted ->*  *Actual* | *Heart Disease = Yes* | *Heart Disease = No* | *Total*  *=* |
| *Heart Disease = Yes* | 123 | 28 | 151 |
| *Heart Disease = No* | 19 | 126 | 145 |
| *Total =* | 142 | 154 | *296* |

The total number of rows used for testing is 296 rows.

To get the precision of the classifier we can use the confusion matrix and use the following formula.

*Precision* = True positives / true positives + false positives

*Precision = 123 / 123 + 19 = 86%*

*Precision is defined as the exactness of the classifier as which values the classifier labeled as positive and which are positive.*

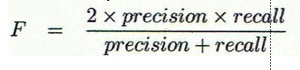
To get a recall of the classifier we can use the confusion matrix and use the following formula.

*Recall = true positives / true positives + false negatives*

*Recall = 123 / 123 + 28 = 81%*

*A recall is defined as what % of positive tuples the classifier labeled as positive.*

Now we can calculate the F score of the classifier by using the precision and recall values of the classifier by using the following formula.



F = 2 X 0.86 X 0.81 / 0.86 + 0.81

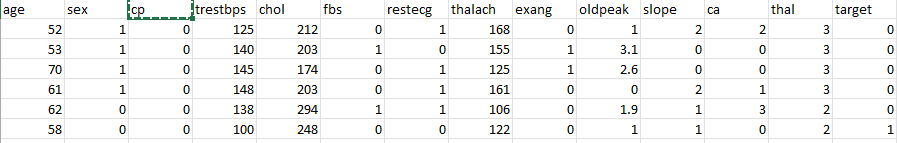
F = 0.8342

Now looking at our F score, accuracy, and other values we have gotten that are all above 80% in this regard. So, we can safely say our **classifier is working with above 80% accuracy.**

# K-Nearest Neighbor:

K nearest neighbors are non-parametric, supervised learning classifiers, which group the data using some type of proximity method such as (Euclidean distance or Manhattan distance) to cluster the data points into individual points.

KNN works best with continuous data contrary to NB. Here we are dropping binary rows or binary-like rows.



Here, we are keeping chol, trestbps, old peak, and thalach to classify the data. Rest are dropped to increase the accuracy of the classifier.



Now we need to divide the data into set and test rows. Here again, we will divide with a 0.225 ratio. Around 22% of the original data set will be used for testing and will be excluded from training.



Now, we need to decide on a value of k. to conclude the value k properly we need to form a graph to check the peak for k.

Chart, line chart

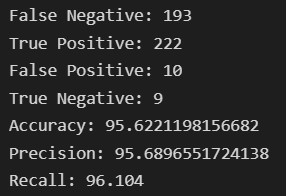
Description automatically generated

From this graph, we can see after a peek of accuracy at 2. Increasing K values is tanking the accuracy on each iteration. So for this classifier to work best we are choosing the value of K at 2!

Now, we are gonna fit and test the data the same as the Naïve Bayes way.

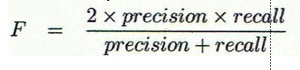
predictions = KNN\_Classifier(X\_train,y\_train,X\_test,y\_test,2)

Now, it’s time to evaluate the classifier using SK learn metrics extension.



|  |  |  |  |
| --- | --- | --- | --- |
| *Predicted ->*  *Actual* | *Heart Disease = Yes* | *Heart Disease = No* | *Total*  *=* |
| *Heart Disease = Yes* | 222 | 9 | 231 |
| *Heart Disease = No* | 10 | 193 | 203 |
| *Total =* | 232 | 202 | *434* |

Now we can calculate the F score of the classifier by using the precision and recall values of the classifier by using the following formula.



F = 2 X 0.95 X 0.95 / 0.95 + 0.95

F = 0.95

Now looking at our F score, accuracy, and other values we have gotten that are all above 95% in this regard. So, we can safely say our **classifier is working with above 95% accuracy.**

**The results are far better than naïve bayes classifier, so KNN is the classifier to go.**

# K-Means Clustering:

K-Means clustering is used to get the graphs from the input values and here we used .describe() function/method on data\_file to get important metrics like mean, standard deviation, minimum value and maximum value. Screenshot is attached below.

A picture containing graphical user interface

Description automatically generated

Here are all the graphs we got using K-Means clustering algorithm.

A picture containing text, crossword puzzle

Description automatically generated

Conclusion:

We can also get some more meaningful and precise insights from this data if we applied artificial neural networks to this data. However, that is beyond the scope of this project. The results we achieved are precise and useful for us to classify whether a given person has a heart condition or not.

References:

* Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial Intelligence, Machine Learning, and Cardiovascular Disease. Clinical Medicine Insights: Cardiology. <https://doi.org/10.1177/1179546820927404>
* Kokol, P., Jurman, J., Bogovič, T., Završnik, T., Završnik, J., & Blažun Vošner, H. (2021). Supporting Real World Decision Making in Coronary Diseases Using Machine Learning. INQUIRY: The Journal of Health Care Organization, Provision, and Financing. <https://doi.org/10.1177/0046958021997338>
* Russak, A. J., Chaudhry, F., De Freitas, J. K., Baron, G., Chaudhry, F. F., Bienstock, S., Paranjpe, I., Vaid, A., Ali, M., Zhao, S., Somani, S., Richter, F., Bawa, T., Levy, P. D., Miotto, R., Nadkarni, G. N., Johnson, K. W., & Glicksberg, B. S. (2020). Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype. Journal of Cardiovascular Pharmacology and Therapeutics, 379–390. <https://doi.org/10.1177/1074248420928651>