Comparing Machine Learning Algorithm to Predict Heart Attacks

Halé Kpetigo
Bethesda, Maryland
hale.kpetigo@gmail.com

***Abstract*—Can heart attacks be predicted? Heart diseases affect many Americans and is the leading cause of death in the US. Sometimes, the condition of a heart disease is silent and not diagnosed until the patient experiences symptoms of heart attacks, arrhythmia or heart failure. If we can diagnose at risk population from safe populations we could save lives.**

The risk factors of heart disease include high blood pressure, high cholesterol, smoking, diabetes, obesity, unhealthy diet, physical inactivity and excessive use of alcohol. We have identified a dataset from University of California, Irvine (UCI) which contains similar dataset. The UCI dataset was developed using patient data from Cleveland, Hungary, Switzerland, and Long Beach, VA. The source database contains 76 attributes however, the published version for research contains 14 attributes which describe the identified risk factors. These attributes include resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, etc…

Using the UCI dataset and unsupervised machine learning algorithms, can we identify patients that belong to a population that is at risk of having a heart attack from a healthy population? We use elbow approach to validate the number of clusters identified by each unsupervised learning algorithm. Comparing our results, we observe that hierarchical clustering provided us with the most accurate clustering – 2 – which corresponds to the number of possible targets: less chance of heart attack vs more chance of heart attack. K-Means and Density clustering identified an optimal number of clusters to be 3.

We then used supervised machine learning algorithms to predict the patient outcome and measured the accuracy of each of these algorithms. The algorithms we retained were Support Vector Machine (SVM), Decision Tree and Gaussian Naïve Bayes. We observed that the three algorithms' accuracy are comparable. They accuracy ranged from 75% to 85% which is significant for predicting patient outcome. SVM performed the best on a larger dataset. We have also observed that SVM accuracy improved on the smaller dataset after being first trained on a larger dataset. Decision Tree offered the most accuracy on the larger dataset and performed the poorest on the smaller dataset. Gaussian Naïve Bayes remained the most consisted across dataset sample size.

From the observed result, heart attacks can be predicted using any of these algorithms. Gaussian Naïve Bayes offers the most reliable accuracy across dataset.

Keywords— machine-learning, support vector machine, ml, k-means, decision-tree hierarchical-clustering, density-based-clustering, Bayesian, Gaussian Naïve Bayes, heart-attack

I. Background

Can heart attacks be predicted? Heart diseases affect many Americans and is the leading cause of death in the US. Sometimes, the condition of a heart disease is silent and not diagnosed until the patient experiences symptoms of heart attacks, arrhythmia or heart failure. If we can diagnose at risk population from safe populations we could save lives.

Machine learning is a tool the is more and more used the the medical arena to predict to outcome. In this paper, we with different machine learning approches to attempt to answer this question.

To answer our question, we first need to identify a reliable dataset to base our analysis one. We identified Heart Disease Data Set from University of Carlifornia, Irvine (UCI). This dataset contained features that could help answer our question. We first analyze this dataset to confirm that it would be a reliable source to answer our question.

Using unsupervised learninng algorightms we attempted to classify the UCI data, to further confirm that classes can be created to group populations that would match the observed outcome in the source data. We compared the efficiency efficientcy of these algorithms in identifiying the different classes of population.

Using supervised learning algorithms, we attempted to predict the outcome of a patient. We measured the effectiveness of each of the algorithms used.

II. Data analysis

A. Identifying heart conditon dataset

The risk factors of heart disease include high blood pressure, high cholesterol, smoking, diabetes, obesity, unhealthy diet, physical inactivity and excessive use of alcohol. We have identified a dataset from University of California, Irvine (UCI) which contains similar dataset. The UCI dataset was developed using patient data from Cleveland, Hungary, Switzerland, and Long Beach, VA. The full dataset contains 1025 records. We also used a smaller sample of the UCI dataset which contains only the 303 records of the Cleveland patients. We used these datasets intechangably to compare how the algorithms performed against them and if there were any observable differences. The source database contains 76 attributes however, the published version for research contains 14 attributes which describe the identified risk factors. These attributes are detailed in Table I, they include resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, etc…

TABLE I. UCI HEART DISEASE DATASET ATTRIBUTE INFORMATION

Features	Description
Age	Age of patient in years
Sex	Sex of patient (1 = male; 0 = female)
cp	Chest pain type (4 values)
trestbps	Resting blood pressure
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar > 120 mg/dl
restecg	Resting electrocardiographic results (values 0,1,2)
thalach	Maximum heart rate achieved
exang	Exercise induced angina
oldpeak	Oldpeak = ST depression induced by exercise relative to rest (ECG)
slope	The slope of the peak exercise ST segment (ECG)
ca	Number of major vessels (0-3) colored by flourosopy
thal	Thalassemia: 0 = normal; 1 = fixed defect; 2 = reversible defect
target	0= less chance of heart attack 1= more chance of heart attack

UCI provides details of the original class distribution of the complete dataset though the Cleveland database is the only one that has been used by Machine Learning researchers to this date. The "goal" field in within the source 76 features, refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). This is important information when discussing the identified clusters using the different unsupervised learning approaches. Table II below further describes the class distribution of the UCI heart disease dataset.

TABLE II. UCI HEART DISEASE DATA SET CLASS DISTRIBUTION

Database	Class Distribution					
	0	*1*	*2*	*3*	*4*	*Total*
Cleveland	165	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

B. *Analysis of the raw data*

Using seaborn pair plot, we first plot the raw data to show any relationship that could exist between the 14 features present in the dataset. “Fig. 1” shows a sample of the pair plot of the studied dataset. The blue color represent the population with less chance of heart attacks while the orange shape shows the population with more chances of heart attack.

We can observe that pairs of the features that have a non continuous value (i.e. sex or cp) the clusters are not globular but instead linear. Each line representing the male or female for sex or each level of chest pain for cp.

From the diagonal distribution plot, we can observe that at risk population has a shorter life expectancy.

We can also observe that our dataset has less women than men, but of the studied women population, women are more at risk compared to the men within their studied population.

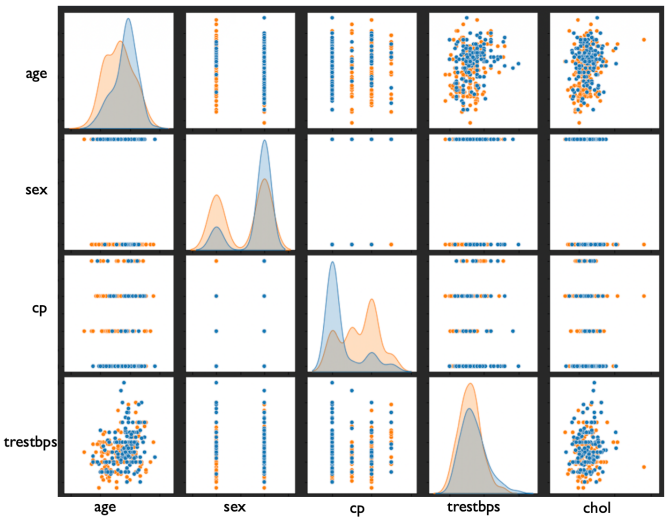

Fig. 1. Sample Pair Plot analysis of UCI heart disease dataset

III. UNSUPERVISED LEARNING

We attempted to classify the populations present in the UCI heart disease dataset using K-Means, Hierachical clusterning, and Density clustering. When possible we used Elbow to identify the optimial number of clusters, and compared with back number of clusters proposed by the implemented algorithm.

A. *K-Means*

Using K-Means, we where able to identify 3 main clusters. “Fig. 2” diplays the ploted 3 clusters. Using Elbow we have are able to confirm that the optimal number of clusters represented in the dataset is 3. This number of cluster fall between the class distibution communicated by UCI. UCI documents that the origininal dataset contains 5 classes. One class representing an absence of a heart attack risk and 4 classes representing 4 different levels of risk.

We also know that the Cleveland data that is used for Machine Learning only distinguishes two targets: at risk and not at risk.

The K-Means Elbow algorithm does show that clusters between 2 and 5 are possible and can be experimented with. To have further insight in the proposed number of cluster 3, we would need to further analyze the data, and understand the science behind it. This may help us justify the existance of the 3 classes. A hypothesis could be that just as UCI proposed 4 level of risks, maybe K-Means is proposing 2 levels of risks.

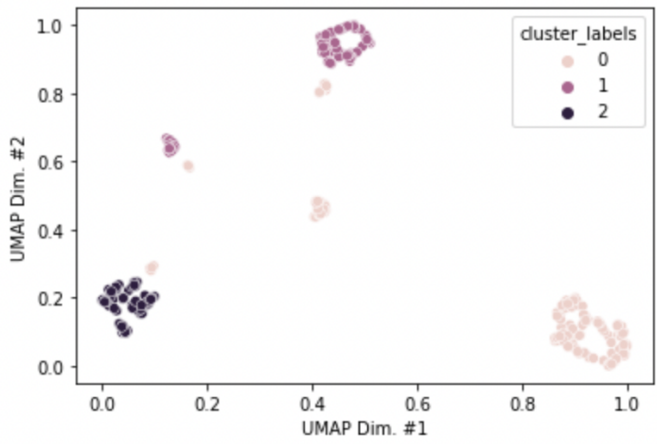

Fig. 2. K-Means plot of reduced UCI heart disease dataset

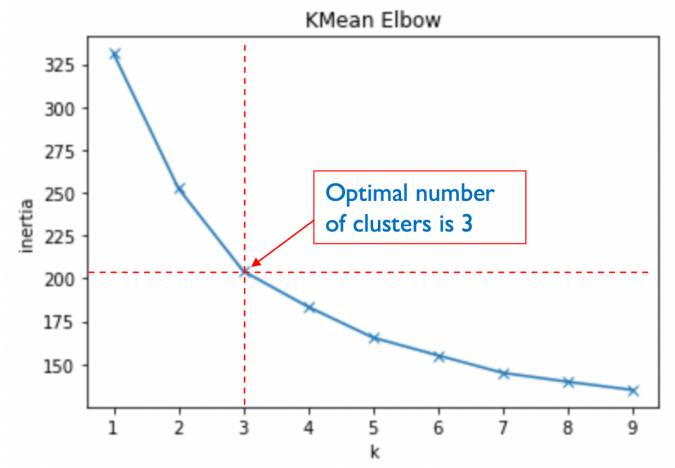

Fig. 3. K-Means Elbow

B. *Hierarchical Clustering*

Using Hierarchical Clustering, we where able to identify 2 main clusters. “Fig. 4” diplays the dendogram which plots the hierarchy between the features of the dataset. It is unclear from what the hierarchy are, however, based on initial observation from the pair plot we could infer that age, sex, or chest pain could be in the predominant features. We will later confirm this in from the class important diagram in the Support Vector Machine approach. Using Elbow in “Fig. 5” we have are able to confirm that the optimal number of clusters represented in the dataset is 2. This corresponds to the two target states present in Cleveland dataset: at risk vs not at risk.

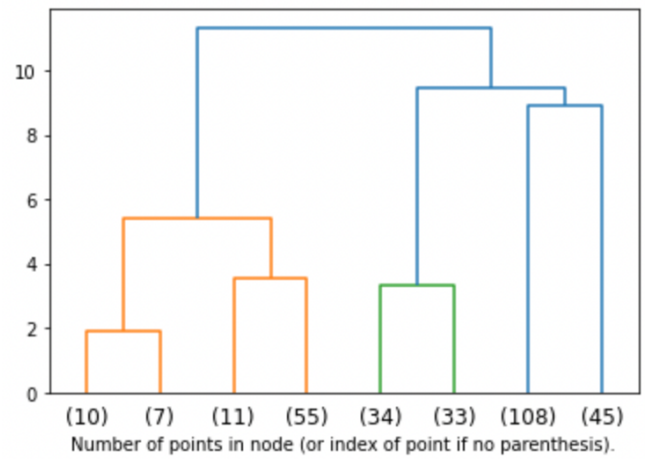

Fig. 4. Dendrogram plot

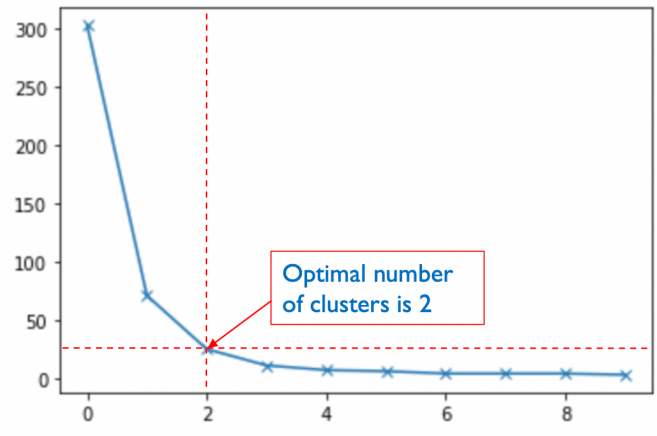

Fig. 5. Hierachical Elbow

C. *Density Clustering*

Similarely, Density clustering also proposed 3 clusters which are plotted in the “Fig. 6”. Here as well, we can observe that the proposed number of cluster in between 2 and 5 which may again be a clasiffication of absence of risk and 2 levels of risks.

The scatter plot of the classes is not visually compelling and difficult to read. It would be interesting to see it this plot improves if we reduice the amount of dimensions before conducting our density clustering analysis.

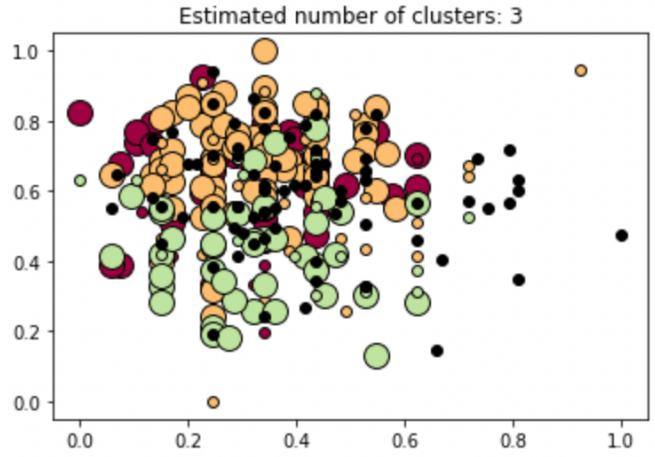

Fig. 6. Density clustering scattered plot

IV. Supervised Learning

In this section, we discuss our implementation of supervised learning algorithms to predict patients outcome. Support Vector Machine, Decision Tree and Gaussian Naïve Bayes are used. For each implementated algorithm we measure the accurary of the algorithm and compares them.

A. *Support Vector Machine*

SVM implementation provided us with a successful prediction of heart attachs with a precision of 81% on the Cleveland dataset and 84% on the complete UCI dataset. We have observed that when we first train the algorithm using the complete UCI dataset, the prior to testing prediction on the Cleveland dataset, we achieve a slighly higher accurary rate at 83% on the Cleveland dataset as Test data.

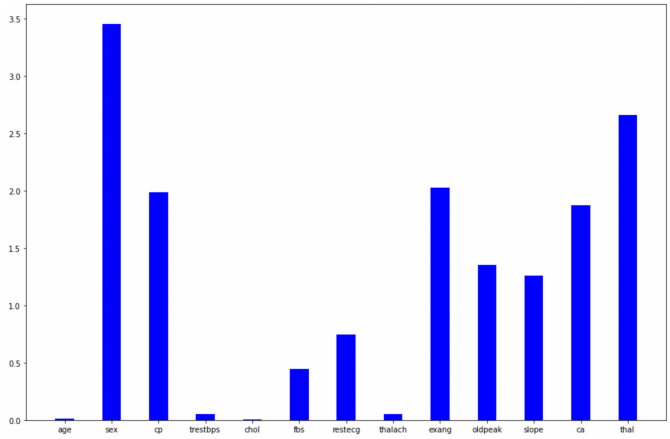

Fig. 7. Feature importance

We have also completed an analysis of the importance of the features that impact the SVM algorithm. We observe that sex, thalassemia, chest pain and exercise induced angina have the highest impact on the algorithm. These features may be the feature present in the Dendrogram of "Fig. 4". At least for sure sex, would be at the top of the hierarchy.

B. *Decision Tree*

We ran a Decision tree algotithm on both the complete UCI dataset and the Cleveland subset. We observed that this algorithm can also successfully predict hearth attacks. The accuracy on the default setting is 77% on the Cleveland dataset. When we set the minimum sample split which describes the number of samples required to split a node to 50, we observe that the accurary is recuded to 75%. When applied to the complete UCI dataset we get respectively 98% and 85% for default settings and 85% for minumum sample split of 50. 98% is by far the highest accuracy of all algorithms.

C. *Gaussian Naïve Bayes*

Finally we implemented Gaussian Naïve Bayes and run our implementation on both the complete UCI dataset and the Cleveland subset. We achieved 80% and 81% accuracy on the Cleveland subset and UCI respectively. Naïve Bayes offered the most consistance accuracy across the datasets.

TABLE III. Supervised Algorithms Accuracy

	Accuracy		
Algorithm	**Cleveland**	**UCI All**	**Transfer**
SVM	81%	84%	83%
Decision Tree	75%	98%	75%
Gaussian Naïve Bayes	80%	81%	80%

V. Conclusion

Heart attacks can be predicted using machine learning algorithms.

Using pair plot, we were able to observe relationships that may exist between the studied datasets. Particularely, we observed that pairs of the features that have a non continuous value (i.e. sex or cp) the clusters are not globular but instead linear. Each line representing the male or female for sex or each level of chest pain for cp.

We observed that at risk population has a shorter life expectancy. We also observed that our dataset has less women than men, but of the studied women population, women are more at risk compared to the men within their studied population.

The different clustering algorithms showed us that clusters between 2 and 5 were possible for grouping the studied population however, the recommended optimal cultering were 2 or 3. These clusting levels would allow us to predict at risk populations.

While all three algorithm studied allowed us to predict patient outcome for heart attacks, Decision Tree algorithm offered us the most accurate result when predicting heart attack risk.

The performance of the 3 algorithms were comparable ranging from 75% to 98%.

SVM performed better on a larger data set.

We have also observed that SVM performance improved on the smaller dataset after being trained on a larger training set

Decision Tree offered the most accuracy on a larger dataset, and performed the poorest on a smaller dataset. This may be due to overfitting. With Decision Tree, we were able to reach an accuracy of 98%.

Gaussian Naïve Bayes remained the most consistent across dataset sample size.