ARTIFICIAL INTELLIGENCE

FINAL REVIEW

"HEART DISEASE PREDICTION SYSTEM"

TEAM MEMBERS

Lade Vivek – 20BCE0432

Mayank Puri – 20BDS0256

Sai Abhinav Kolli – 20BCE2272

Virupakshi Dharaneswar Reddy – 20BCE2351

Ahrar Karim – 20BCI0311

SUBMITTED TO

Prof. Mohana C M

CONTENTS:

Abstract

The system for detecting cardiac illness utilising artificial intelligence and machine learning algorithms is the main topic of the study. We demonstrate how artificial intelligence can be used to forecast if someone would get cardiac disease. A python-based application is created for healthcare research in this study since it is more dependable and aids in tracking and establishing various kinds of health monitoring applications. We demonstrate categorical variable manipulation and the conversion of categorical columns in data processing. We outline the key stages of application development, including the gathering of databases, applying logistic regression, and assessing the features of the dataset. A random forest classifier method is created to detect cardiac problems more accurately. This application, which is deemed important based on its about 83% accuracy rate across training data, uses data analysis. The K-Means Clustering algorithm, including the tests and findings, is then covered. This algorithm improves the accuracy of study diagnosis. The paper's goals, constraints, and research contributions are presented towards the end.

Introduction

Clinical decision support systems frequently employ artificial intelligence approaches for accurate disease prediction and diagnosis. Due to their capacity to uncover hidden patterns and connections in the medical data supplied by medical practitioners, these classification algorithms are extremely useful when creating clinical support systems. Given that heart disease is one of the top causes of mortality worldwide, diagnosing heart disease is one of the most crucial purposes for such systems. Nearly all systems that foretell heart illnesses using clinical datasets include complicated laboratory tests as inputs and parameters. None of the algorithms can forecast cardiac illnesses based on risk factors such age, family history, diabetes, high blood pressure, high cholesterol, smoking, drinking, obesity, and inactivity, among others. Many of the obvious risk indicators that might be utilized to diagnose heart disease are shared by patients. A

system based on such risk factors would benefit medical professionals as well as patients by alerting them to the possibility of heart disease before they enter a hospital or undergo pricey diagnostic tests.

Motivation

The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Additionally, the goal of this research is to determine the optimum classification method for detecting cardiac disease in a patient. Three classification algorithms, namely Naive Bayes, Decision Tree, and Random Forest, are employed at various levels of evaluations in a comparative study and analysis to support this work. Although these machine learning methods are widely utilised, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to evaluate the three algorithms. Researchers and medical professionals will be able to better understand the situation thanks to this.

Objectives

1. The goal of our heart disease prediction project is to determine if a patient should be diagnosed with heart disease or not, which is a binary outcome so:
    - Positive result = 1, the patient will be diagnosed with heart disease.
    - Negative result = 0, the patient will not be diagnosed with heart disease.
2. We must find which classification model has the greatest accuracy and identify correlations in our data. Finally, we also must determine which features are the most influential in our heart disease diagnosis.

Literature Survey

Research Paper-1

Methodology	Advantages	Limitations	Accuracy
This paper shows that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. We will evaluate the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier.	1. Shows the risk of heart attack based on their age and resting blood pressure. 2. Shows the analysis whether the patient is having heart attack or not based on sex and type of chest pain.	1. We can also use different algorithms like SVC, Decision tree algorithm but it won't produce better accuracy. 2. This method is cost-efficient (saves money) and faster than other algorithms.	The patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and KNN to get an accuracy of an average of 87.5% on our model which is better than the previous models having an accuracy of 85%. Also, it is concluded that accuracy of KNN is highest between the three algorithms that we have used i.e., 88.52%.

Research Paper-2

Methodology	Advantages	Limitations	Accuracy
The algorithms used in this research are Naive Bayes, Decision Tree, Random Forest, Logistic Regression. The dataset used was the Heart disease Dataset, which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease.	1. Attempt to detect these heart diseases at early stage to avoid disastrous consequences. 2. Pre-processed UCI dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. 3. Predicting the best algorithm with good accuracy	1. Data which is stored with in the dataset should be accurate to real world because we will be training the model as per the values in it. 2. This method is cost-efficient.	

Algorithm	Accuracy
Decision Tree	81.97%
Random Forest	90.16%
Naive Bayes	85.25%
Logistic Regression	85.25%

Research Paper-3

<table>
<tr><th>Methodology</th><th>Advantages</th><th>Limitations</th><th>Accuracy</th></tr>
<tr>
<td>The authors have implemented several machine learning algorithms, Logistic Regression has given an accuracy of 93%, Random Forest 92% and Gaussian Naive Bayes 90%, we gave notice that the results are close with simple progression of Logistic Regression. We have tested the diagnosis of heart patients by applying two techniques: genetic algorithms and the KNN algorithm. The results gave satisfaction with the KNN algorithm.</td>
<td>1. From different algorithms we have used so far, we have got better accuracy and stability for Neural networks, KNN, and SVM.

2. Here, in this method we can select the best algorithm suitable for getting accuracy and results for this system.</td>
<td>1. authors have used artificial neural networks now unfortunately, they are not able to show their results.

2. The Authors did not mention their results with the usage of algorithms.</td>
<td>
<table>
<tr><th>Algorithm Used</th><th>Authors</th><th>Accuracy</th></tr>
<tr><td>Neural Networks</td><td>T John Peter et al., 2012
Chaitrali S et al., 2012</td><td>78%
100%</td></tr>
<tr><td>KNN</td><td>T John Peter et al., 2012

C.Kalaiselvi 2016</td><td>75%
87%</td></tr>
<tr><td>SVM</td><td>T John Peter et al., 2012
Chaitrali S et al., 2012
Shamsher Bahadur et al., 2013
B.Venkata-lakshmi et al., 2014</td><td>76%
99%
99%

84%</td></tr>
</table>
</td>
</tr>
</table>

Research Paper-4

Methodology	Advantages	Limitations	Accuracy
The data for 50 people was collected from surveys done by the American Heart Association. Most of heart disease patients had many similarities in the risk factors. Data analysis has been carried out in order to transform data into a useful form, for this the values were encoded mostly between a range [-1, 1]. Data analysis also removed the inconsistency and anomalies in the data. This was needed. Data analysis was needed for correct data preprocessing. The removal of missing and incorrect inputs will help the neural network to generalize well. we have implemented AI algorithms such as 1. K Neighbours Classifier 2. Navie Bayies 3. Decision Tree Classifier 4. Random Forest Classifier	Need more datasets, to increase the accuracy of the algorithms. The proposed application can only be used by medical Personnel. The proposed application is Web-based, hence can not be used on Mobile devices. The result of the application depends upon the accuracy of the algorithms.	User can search for doctor's help at any point of time. User can talk about their Heart Disease and get instant diagnosis. Doctors get more clients online. Very useful in case of emergency.	The above table 1.1show that the best accuracy on the given dataset is 88% and The lowest accuracy of 78%. The Naïve Bayes has the highest accuracy while the Decision A tree has the lowest accuracy. By observing other performance measures that are used for results too. TP rate of KNN is 40, ANN is 36, Naïve Bayes is 36, The decision tree is 29, and the Random forest is 33. This shows that the KNN has the highest TP rate and Decision Tree has the lowest TP rate. Similarly, the Decision tree has the highest FP rate of 15 and the ANN has the lowest FP rate of 2.

Research Paper-5

<table>
<tr><th>Methodology</th><th>Advantages</th><th>Limitations</th><th>Accuracy</th></tr>
<tr>
<td>I will be using the experimental type of research design. Itquantitative research method. Basically, it is research conducted with a a scientific approach, where a set of variables are kept constant while another set of variables e being measured as the subject of the experiment. This is more practical while conducting face e recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross-checked with revious data..</td>
<td>To enhance visualization and ease of interpretation.

Extensive experiments on real-world large datasets have demonstrated the effectiveness of our approach for prediction of heart disease

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

Reduce the cost of medical tests.</td>
<td>Medical diagnosis is considered as a significant yet intricate task that needs to be carried out

precisely and efficiently. The automation of the same would be highly beneficial. Clinical

decisions are often made based on doctor’s intuition and experience rather than on the

knowledge rich data hidden in the database.
.</td>
<td>
<table>
<tr><th>Algorithm</th><th>Accuracy</th></tr>
<tr><td>Decision Tree</td><td>74.97%</td></tr>
<tr><td>Random Forest</td><td>85.16%</td></tr>
<tr><td>Naive Bayes</td><td>85.25%</td></tr>
<tr><td>k-nearest</td><td>68.57%</td></tr>
<tr><td>Logistic Regression</td><td>74.25%</td></tr>
</table>
</td>
</tr>
</table>

Research Paper-6

<table>
<tr><th>Methodology</th><th>Advantages</th><th>Limitations</th><th>Accuracy</th></tr>
<tr>
<td>The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.
· 1.Collection of dataset : Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data.
· 2. Selection of attributes : Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction.
· 3. Pre-processing of Data : Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can</td>
<td>Cardiovascular disease prediction aids practitioners in making more accurate health decisions for their patients. Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care.

Artificial inteligence (AI) is a plausible option for reducing and understanding heart symptoms of disease.</td>
<td>The main motivation for doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient</td>
<td>
<table>
<tr><th>Algorithm</th><th>Accuracy</th></tr>
<tr><td>Decision Tree</td><td>75.97 %</td></tr>
<tr><td>Random Forest</td><td>79.16 %</td></tr>
<tr><td>Naive Bayes</td><td>76.25 %</td></tr>
<tr><td>Adaboost</td><td>73.57 %</td></tr>
<tr><td>Logistic Regression</td><td>74.25 %</td></tr>
<tr><td>XG BOOST</td><td>81.3 %</td></tr>
</table>
</td>
</tr>
</table>

cause misleading outcomes. · **4. Balancing of Data :** Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling. **5. Prediction of Disease:** Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification			

Research Paper 7

Methodology	Advantages	Limitations	Accuracy
KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time. A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. For example, in heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk	Cardiovascular disease prediction aids practitioners in **making more accurate health decisions for their patients.** Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Artificial intelligence (AI) is a plausible option for reducing and understanding heart symptoms of disease.	Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient	By Using K-Nearest Neighbour Algorithm, we achieved 84.44% correct classification rate and 15.56% min classification rate.

minimization			

Research Paper 8:

Methodology	Advantages	Limitations	Accuracy
K Nearest Neighbor (KNN): It is preferred when parameters are continuous. In KNN, classification is done by predicting the nearest neighbor. It is preferred over other classification algorithms due to its simplicity and high speed. It can be used to solve both classification and regression problems. The algorithm takes the heart disease data set and classifies whether a person has heart disease or not. KNN captures the idea by calculating the distance between points on a graph. **Naive Bayes**:Naive Bayes is used for a classification based on Bayes' theorem. Occurrences of Particular characteristics of a class are independent of the presence or absence of other characteristics according to the	To enhance visualization and ease of interpretation. · Extensive experiments on real-world large datasets have demonstrated the effectiveness of our approach for prediction of heart disease	Need more datasets, to increase the accuracy of the algorithms. · The proposed application can only be used by Medical Personnel. The proposed application is Web-based, hence cannot be used in Mobile devices	KNN - 0.87 ANN - 0.87 Naive Bayes - 0.88 Decision Tree - 0.78 Random Forest - 0.82

naive Bayesian classifier theorem			

Research Paper 9:

Methodology	Advantages	Limitations	Accuracy
Preprocess Dataset: DataPreprocessing is\a technique that is used to convert raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis Supervised Learning Algorithms: Decision Tree: Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes,	Cardiovascular disease prediction aids practitioners in **making more accurate health decisions for their patients.** Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Artificial intelligence (AI) is a plausible option for this	Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. This practice leads to unwanted biases, errors and excessive medical costs which affect the quality of service provided to patients.	KNN - 0.87 ANN - 0.87 Naive Bayes - 0.88 Decision Tree - 0.78 Random Forest - 0.82

and the outer branches are the outcome. Decision Trees are chosen because they are fast, reliable, easy to interpret, and very little data preparation is required			

Research Paper 10:

Methodology	Advantages	Limitations	Accuracy
In the dataset the "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The dataset does not have any null values but many outliers needed to be handled properly, and also the dataset is not properly distributed. Various plotting techniques were used for checking the skewness of the data, outlier detection, and the distribution of the data. All these preprocessing techniques play an important role when passing the data for classification or prediction purposes. Machine Learning Classifiers Proposed approach was applied to the dataset in which	Description of the Dataset. dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes. The computational time was reduced which is helpful when deploying the models.	The dataset is not properly distributed. By analyzing the distribution plots, it is visible that thal and fasting blood sugar is not uniformly distributed and they needed to be handled; otherwise, it will result in overfitting or underfitting of the data. The dataset is not normalized, there is no equal distribution of the target class, it can further be seen when a correlation heatmap is plotted, and there are so many negative values.	Random Forest 76.7% Logistic Regression 83.64% KNeighbors 82.27% Support Vector Machine 84.09% Decision Tree 75.0% XGBoost 70.0%.

firstly the dataset was properly analyzed and then different machine learning algorithms consisting of linear model selection in which Logistic Regression was used. For focusing on neighbour selection technique KNeighborsClassifier was used, then tree-based technique like DecisionTreeClassifier was used, and then a very popular and most popular technique of ensemble methods RandomForestClassifier by using the cross-validation and grid search for finding the best parameters or in other words doing the hyperparameter tuning.			

Research Paper 11:

Methodology	Advantages	Limitations	Accuracy
(i) Dataset of training (ii) Dataset of testing (iii) Checking the shape/features of the input (iv) The procedure of initiating the sequential layer (v) Adding dense layers with dropout layers and ReLU activation functions (vi) Adding a last dense layer with one output and binary activation function (vii) End repeat (viii) L (output) (ix) End procedure	Better predictive accuracy than filter methods. It renders good feature subsets for the used algorithm. And then for selecting the selected features, select from the model which is a part of feature selection in the scikit-learn library.	There might be a chance if duplicates are not dealt with properly; they might show up in the test dataset which is also in the training dataset.	Random Forest 88% Logistic Regression 85.9% KNeighbors 79.69% Support Vector Machine 84.26% Decision Tree 76.35% XGBoost 71.1%

Research Paper 12:

Methodology	Advantages	Limitations	Accuracy
A sequential model with a fully connected dense layer is used, with the flatten and dropout layers to prevent the overfitting and the results are compared of the machine learning and deep learning and variations in the learning including computational time and accuracy can be analysed Then for checking how well a model is performing, an accuracy score is used. It is defined as	Accuracy is greater than the machine learning models. Algorithm produced greater accuracy and more promising than other approaches. The comparison of different classifiers of ML and DL. The whole knowledge which will be obtained could be transferred to the mobile devices means, when the person will input these symptoms in the mobile device in which the trained model will already be	It was also found out that the statistical analysis is also important when a dataset is analyzed and it should have a Gaussian distribution. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well.	Random Forest 80.3% Logistic Regression 83.31% KNeighbors 84.86% Support Vector Machine 83.29% Decision Tree 82.33% XGBoost 71.4%.

the true positive values plus true negative values divided by true positive plus true negative plus false positive plus false negative. the formula is Accuracy = (TP + TN)/ (TP + TN + FP + FN) The dataset is normalized, and the feature selection is done and also the outliers are handled using the Isolation Forest.	present and then can analyze the symptoms and could give the prescription accordingly. Different doctors could be taken under consideration and a complete autonomous system could be generated.		

Research Paper- 13

<table>
<tr><th>Methodology</th><th>Advantages</th><th>Limitations</th><th>Accuracy</th></tr>
<tr>
<td>I will be using the experimental type of research design. where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data.</td>
<td>Shows the analysis whether the patient is having heart attack or not based on sex and type of chest pain.

Predicting the best algorithm with good accuracy

To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.

Low cost for testing</td>
<td>Medical diagnosis is considered as a significant yet intricate task that needs to be carried out

precisely and efficiently. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. (Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.)</td>
<td>
<table>
<tr><th>Algorithm</th><th>Accuracy</th></tr>
<tr><td>Random Forest</td><td>79.16 %</td></tr>
<tr><td>Naive Bayes</td><td>88.25 %</td></tr>
<tr><td>KNN</td><td>93.57 %</td></tr>
</table>
</td>
</tr>
</table>

Research Paper 14:

Methodology	Advantages	Limitations
Random Forest: Both classification and regression are done using Random Forest algorithms. The data is organised into a tree, and predictions are based on that tree. Even with a substantial number of record values missing, the Random Forest algorithm can still produce the same results when applied to huge datasets. The decision tree's generated samples can be preserved and used to different sets of data. In random forest, there are two stages: first, generate a random forest, and then, using a classifier produced in the first stage, make a prediction. ii. Decision Tree: A flowchart-like representation of the decision tree algorithm, where the outer branches reflect the results, and the inside nodes represent the dataset properties. Decision trees are used because they are quick, dependable, simple to understand, and require very little data preparation. In a decision tree, the class label prediction comes from the tree's root. The root attribute's value is contrasted with the record's attribute. The matching branch is followed to the value indicated by the comparison result, and a jump is then made to the following node. iii.KNN: The k-nearest neighbours (KNN) algorithm is a supervised machine learning technique that may be applied to classification and regression predicting issues. However, it is primarily employed in industry for classification and forecasting problems. The next two characteristics would accurately describe KNN: KNN is a lazy learning algorithm since it uses all of the data for training while classifying rather than having a separate training phase. KNN is another example of a non-parametric learning algorithm because it makes no assumptions about the underlying data.	Cardiovascular disease prediction aids practitioners in **making more accurate health decisions for their patients.** Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Artificial intelligence (AI) is a plausible option for reducing and understanding heart symptoms of disease.	Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. It was also found out that the statistical analysis is also important when a dataset is analyzed and it should have a Gaussian distribution. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well.

Table I VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHMS

Algorithm	True Positive	False Positive	False Negative	True Negative
KNN	839	120	4	496
Decision Tree	915	44	0	500
Random Forest	947	12	0	500

Table II ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS

Algorithm	Precision	Recall	F - measure	Accuracy
KNN	0.92	0.90	0.90	90.15%
Decision Tree	0.97	0.97	0.97	96.25%
Random Forest	0.99	0.99	0.99	99.17%

Research Paper-15

Methodology	Advantages	Limitations
We have used four types of algorithms i.e., Linear Regression, Logistic regression, Svm, Reinforcement Learning. Linear Regression: The cost function is a function that describes the relationship between the data and the way it is related to the real world. The function can be defined as exactly as possible by a straight line, and the linear regression method is one of the best models for this purpose. Logistic Regression: Common nonlinear probabilistic regression models like the logistic regression model can only be either 0 or 1. The assumption that p are independent variables X = [x1, x2,.,.] indicates that there is a probability that Y is similar to 1. xp], p = P(y = 1|X). a path to follow. Svm: The C error penalty coefficient, the gamma kernel function, and the kernel function are important parameters in complex data analysis. One of the standard classifications is SVM (Support Vector Machine) which has been replaced by alternative approaches. Reinforcement Learning: The concept of learning bagging in an ensemble is the foundation of the reinforcement education paradigm. A complicated forest is created by combining numerous small decision trees for the most accurate forecast.	1. Attempt to detect these heart diseases at early stage to avoid disastrous consequences. 2. Here, in this method we can select the best algorithm suitable for getting accuracy and results for this system. Shows the analysis whether the patient is having heart attack or not based on sex and type of chest pain.	1. This method is cost-efficient (saves money) and faster than other algorithms. 2. The proposed application is Web-based, hence cannot be used in Mobile devices.

Accuracy

Methodology	Accuracy	Precision	Recall	F-Measure	Execution time
Linear Regression	0.741	0.785	0.695	0.654	0.52
Logistic regression	0.713	0.801	0.698	0.661	0.34
Svm (support vector machine)	0.699	0.791	0.711	0.713	0.23
Reinforcement Learning	0.805	0.813	0.724	0.743	0.18

Methodology

"K-means Clustering algorithm."

Step 1: Select the Number of Clusters, k

The number of clusters we want to identify is the k in k-means clustering. In this case, since we assumed that there are 3 clusters, k = 3.

Step 2: Select k Points at Random

We start the process of finding clusters by selecting 3 random points (not necessarily our data points). These points will now act as centroids, or the center, of clusters that we are going to make

Step 3: Make k Clusters

To make the clusters, we start by measuring the distance from each data point to each of the 3 centroids. And we assign the points to the cluster closest to it. So, for a sample point

Step 4: Compute New Centroid of Each Cluster

Now that we have our 3 clusters, we find the new centroids formed by each of them.

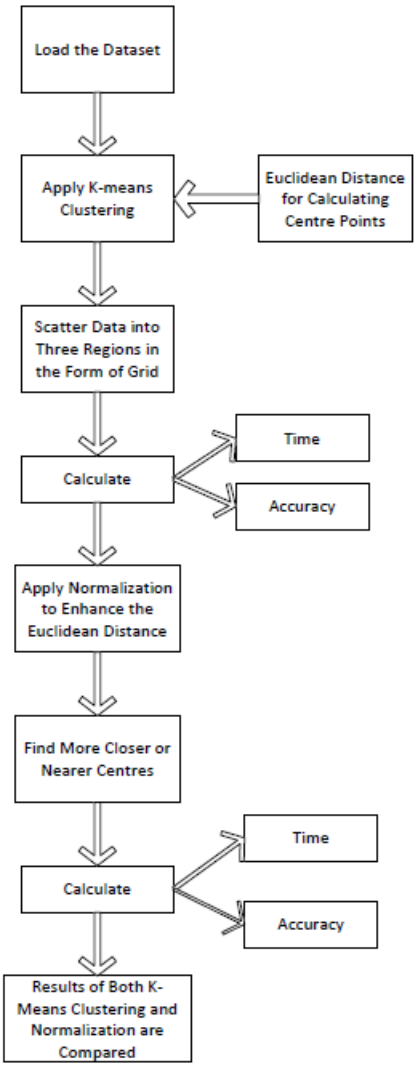
Step 5: Assess the Quality of Each Cluster

Since k-means can't see the clustering as we can, it measures the quality by finding the variation within all the clusters. The basic idea behind k-means clustering is defining clusters so that the within-cluster variation is minimized. We calculate something called Within-Cluster Sum of Squares (WCSS) to quantify this variance

Step 6: Repeat Steps 3–5

Once we have previous clusters and the variation stored, we start all over. But only this time we use the centroids we calculated previously to - make 3 new clusters, recalculate the center of the new clusters, and calculate the sum of the variation within all the clusters.

Architecture



Implementation

In [2]:
```
import numpy as np
from numpy.linalg import norm

class Kmeans:
    '''Implementing Kmeans algo'''

    def __init__(self, n_clusters, max_iter=100, random_state=123):
        self.n_clusters = n_clusters
        self.max_iter = max_iter
        self.random_state = random_state

    def first_centroids(self, X):
        np.random.RandomState(self.random_state)
        random_id = np.random.permutation(X.shape[0])
        centroids = X[random_id[:self.n_clusters]]
        return centroids

    def compute_centroids(self, X, labels):
        centroids = np.zeros((self.n_clusters, X.shape[1]))
        for k in range(self.n_clusters):
            centroids[k, :] = np.mean(X[labels == k, :], axis=0)
        return centroids

    def compute_distance(self, X, centroids):
        distance = np.zeros((X.shape[0], self.n_clusters))
        for k in range(self.n_clusters):
            row_norm = norm(X - centroids[k, :], axis=1)
            distance[:, k] = np.square(row_norm)
        return distance

    def closest_cluster(self, distance):
        return np.argmin(distance, axis=1)

    def compute_sse(self, X, labels, centroids):
        distance = np.zeros(X.shape[0])
        for k in range(self.n_clusters):
            distance[labels == k] = norm(X[labels == k] - centroids[k], axis=1)
        return np.sum(np.square(distance))

    def fit(self, X):
        self.centroids = self.first_centroids(X)
        for i in range(self.max_iter):
            old_centroids = self.centroids
            distance = self.compute_distance(X, old_centroids)
            self.labels = self.closest_cluster(distance)
            self.centroids = self.compute_centroids(X, self.labels)
            if np.all(old_centroids == self.centroids):
                break
        self.error = self.compute_sse(X, self.labels, self.centroids)

    def predict(self, X):
        old_centroids = self.centroids#to define old within function
        distance = self.compute_distance(X, old_centroids)
        return self.closest_cluster(distance)
```

In [3]:
```
# Import the data as well as pandas module
import pandas as pd
df = pd.read_csv("C:\\Users\\ACER\\OneDrive\\Desktop\\VIT Downloads\\healthcare-dataset-stroke-data.csv")
df.head(5)
df.replace(to_replace="formerly smoked",
           value="smokes")
```

Out[3]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	smokes	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	smokes	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

In [4]:
```
#Dealing with the categorical data. Among many methods possible we are using One Hot Encoding here.
a = pd.get_dummies(df['gender'], prefix = "gender")
b = pd.get_dummies(df['ever_married'], prefix = "ever_married")
c = pd.get_dummies(df['work_type'], prefix = "work_type")
d = pd.get_dummies(df['Residence_type'], prefix = "Residence_type")
e = pd.get_dummies(df['smoking_status'], prefix = "smoking_status")

frames = [df, a, b, c, d, e]
df = pd.concat(frames, axis = 1)

df_copy = df.drop(columns = ['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status'])
df_copy.head()
```

Out[4]:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke	gender_Female	gender_Male	gender_Other	...	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	Residence_type_Rural	Residence_type_Urban	smoking_status_Unknown	smoking_status_
0	9046	67.0	0	1	228.69	36.6	1	0	1	0	...	0	1	0	0	0	1	0	
1	51676	61.0	0	0	202.21	NaN	1	1	0	0	...	0	0	1	0	1	0	0	
2	31112	80.0	0	1	105.92	32.5	1	0	1	0	...	0	1	0	0	1	0	0	
3	60182	49.0	0	0	171.23	34.4	1	1	0	0	...	0	1	0	0	0	1	0	
4	1665	79.0	1	0	174.12	24.0	1	1	0	0	...	0	0	1	0	1	0	0	

5 rows × 23 columns

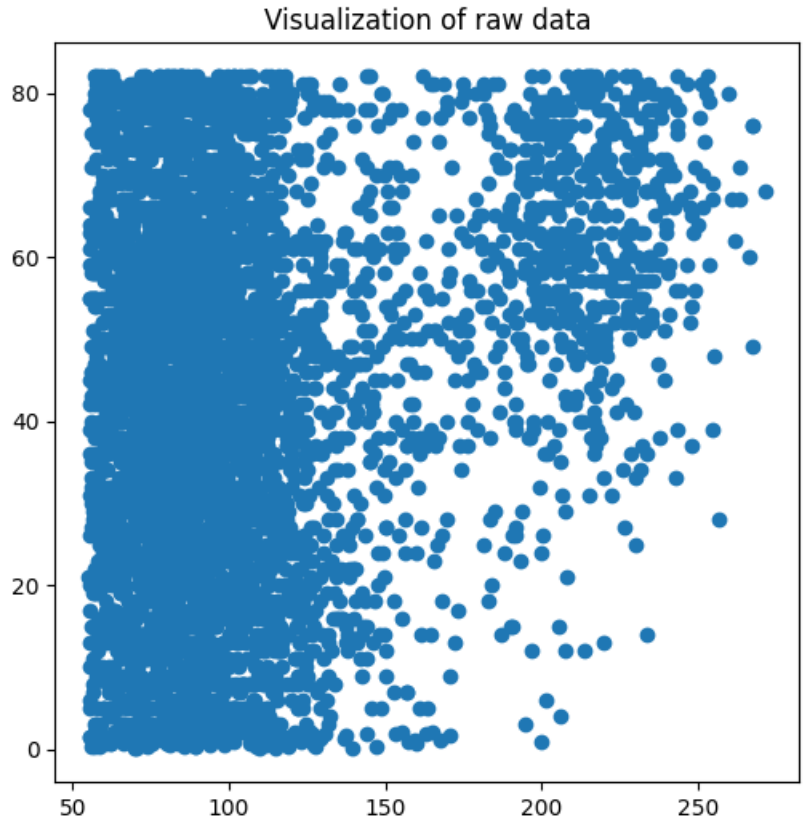
In [5]:
```
#very important to drop the target if it is present
df_drop = df.drop(columns = ['stroke'])
df_drop.head(5)
```

Out[5]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	...	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	Residence_type_Rural	Residence_type_Urban	smoking_status_Unknown	smoking_status_fo
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	...	0	1	0	0	0	1	0	
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	...	0	0	1	0	1	0	0	
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	...	0	1	0	0	1	0	0	
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	...	0	1	0	0	0	1	0	
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	...	0	0	1	0	1	0	0	

5 rows × 27 columns

In [6]:
```
# Plot the data
# Importing some visualizing packages
import matplotlib.pyplot as plt
plt.figure(figsize=(6, 6))
features = ['gender',   'age',  'hypertension', 'heart_disease',        'ever_married', 'work_
X = df_copy['avg_glucose_level']
Y = df_copy['age']
plt.scatter(X, Y)
plt.xlabel('')
plt.ylabel('')
plt.title('Visualization of raw data');
```



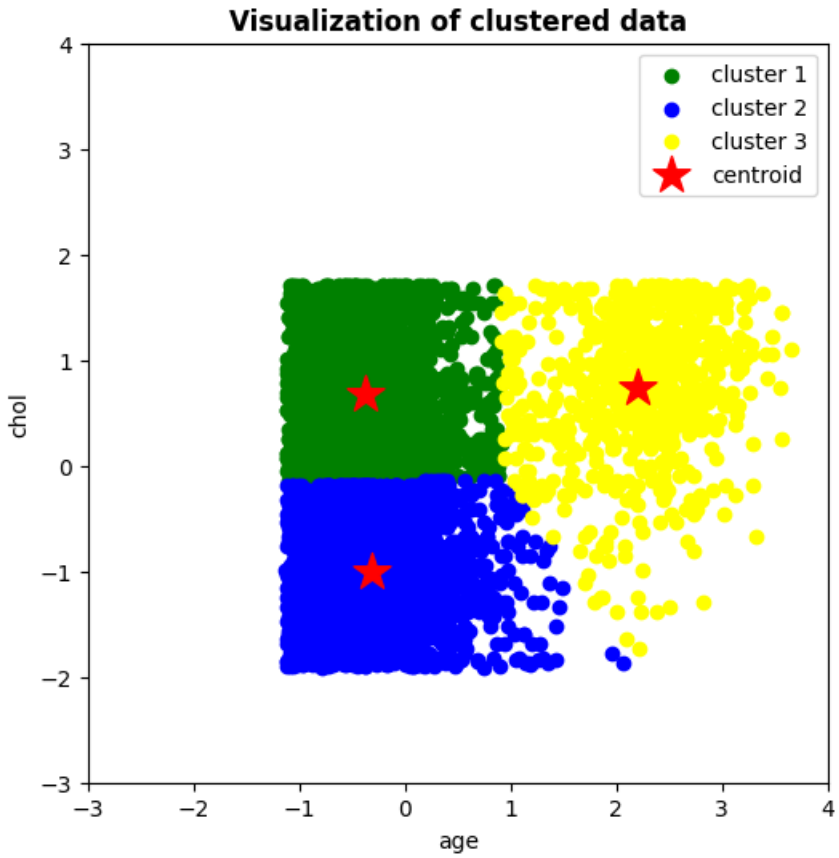
In [7]:
```
#use a subset of the data to start k-means exploration
data = df_copy[['avg_glucose_level', 'age']]
```

In [8]:
```
# Standardize the data
from sklearn.preprocessing import StandardScaler
X_std = StandardScaler().fit_transform(data)

# Run local implementation of kmeans Here we tested 3 clusters
km = Kmeans(n_clusters=3, max_iter=100, random_state = 42)
km.fit(X_std)
centroids = km.centroids
# labels_ are equivalent to calling fit(x) then predict
labels_ = km.predict(X_std)
labels_
```

Out[8]:
```
array([2, 2, 0, ..., 1, 2, 0], dtype=int64)
```

In [9]:
```
#Plotting the clustered data
fig, ax = plt.subplots(figsize=(6, 6))
plt.scatter(X_std[labels_ == 0, 0], X_std[labels_ == 0, 1],
            c='green', label='cluster 1')
plt.scatter(X_std[labels_ == 1, 0], X_std[labels_ == 1, 1],
            c='blue', label='cluster 2')
plt.scatter(X_std[labels_ == 2, 0], X_std[labels_ == 2, 1],
            c='yellow', label='cluster 3')
plt.scatter(centroids[:, 0], centroids[:, 1], marker='*', s=300,
            c='r', label='centroid')
plt.legend()
plt.xlim([-3, 4])
plt.ylim([-3, 4])
plt.xlabel('age')
plt.ylabel('chol')
plt.title('Visualization of clustered data', fontweight='bold')
ax.set_aspect('equal');
```

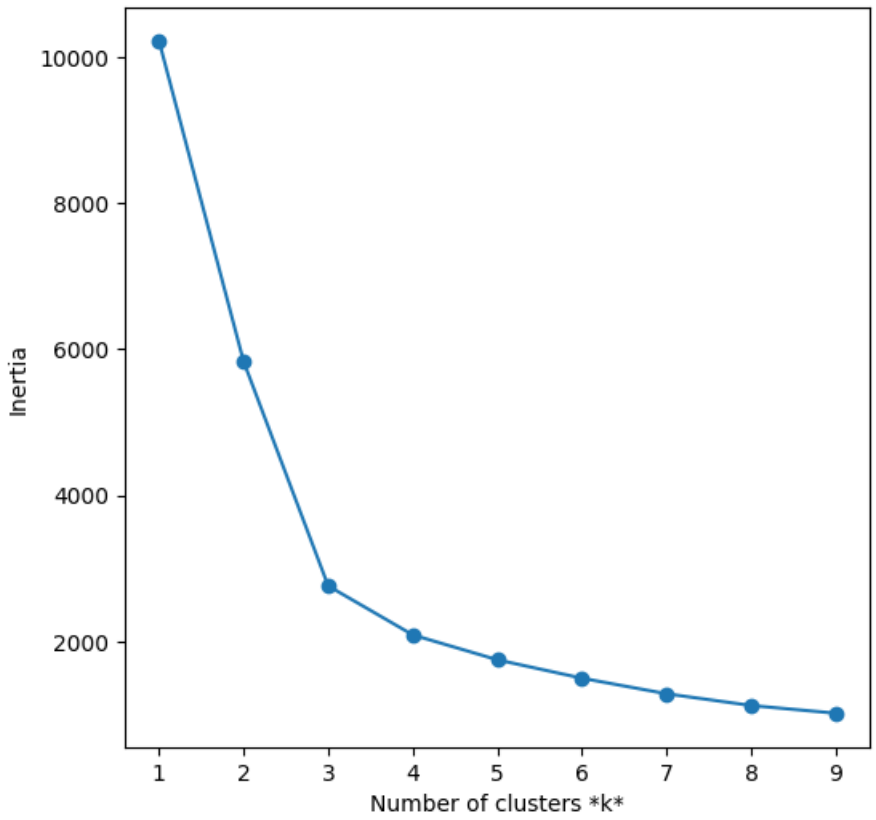


In [11]:
```
#labels added to dataset
data['cluster'] = labels_
data.head(5)
#uses labels from clusters to see on data
fig, ax = plt.subplots()
colors = {0:'red', 1:'blue', 2:'yellow'}
grouped = data.groupby('cluster')
for key, group in grouped:
    group.plot(ax=ax, kind='scatter', x='age', y='chol', label=key, color=colors[key])
plt.show()
```

In [12]:
```
#elbow method:
# Run the Kmeans algorithm and get the index of data points clusters
from sklearn.cluster import KMeans
sse = []
list_k = list(range(1, 10))

for k in list_k:
    km = KMeans(n_clusters=k)
    km.fit(X_std)
    sse.append(km.inertia_)

# Plot sse against k
plt.figure(figsize=(6, 6))
plt.plot(list_k, sse, '-o')
plt.xlabel(r'Number of clusters *k*')
plt.ylabel('Inertia');
```



Conclusion

Furthermore, data standardization, such as derivation standardization, is a useful way for improving performance, such as accuracy. Now, the doctor performs this operation manually based on the waveform's properties. A comprehensive system can be developed by combining the method used in this work with additional research. We contrast K-means clustering algorithm with Reinforcement Learning for prediction of heart stroke. For the experimentation study we used congenital heart disease (CHD) Datasets. According to this comparison analysis, the Reinforcement Learning model outperforms better than the other classifier in terms of accuracy and computation time as 0.805 on CHD dataset. Simultaneously, the results provide a significant reference for clinical data-based diagnosis and the enhancement of medical efficiency.

References:

Heart Disease Prediction using Artificial Intelligence – IJERT

[PDF] Heart Disease Prediction System (researchgate.net)

https://www.researchgate.net/publication/331589020_Heart_Disease_Prediction_System

IJCRT1813083.pdf

HeartDiseasePrediction.pdf

Heart Disease Prediction Using Machine Learning (ijraset.com)

https://www.researchgate.net/publication/349470771_Using_Machine_Learning_for_Heart_Disease_Prediction

https://www.ijert.org/research/heart-disease-prediction-using-machine-learning-IJERTV9IS040614.pdf

https://www.researchgate.net/publication/348604625_Heart_disease_prediction_using_machine_learning_algorithms

https://www.researchgate.net/publication/319486202_PREDICTION_OF_HEART_DISEASE_USING_K-MEANS_and_ARTIFICIAL_NEURAL_NETWORK_as_HYBRID_APPROACH_to_IMPROVE_ACCURACY