PROJECT REPORT

Heart Failure Prediction Analysis

Course: 19CSE304 Foundation of Data Science
Project by: Group Number 3

Team Members	
Adithya Nair	AM.EN.U4CSE19103
Akshay Hari	AM.EN.U4CSE19104
Alekh Avinash	AM.EN.U4CSE19105
Kumar Anurag	AM.EN.U4CSE19130
Sreejith Kumara Pai	AM.EN.U4CSE19153

TABLE OF CONTENTS

- Abstract …… 3
- Introduction ……..…..…… 3
- Broad Context ……4
- Study System ……..... 5
- Methods ……... 11
- Results ……..13
- Discussion ……14
- Conclusion ……....14

Abstract. The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results & making effective decisions on data, some advanced data mining techniques are used. In this study, multiple Heart Disease Prediction models were developed using several machine learning algorithms including KNN, Logistic Regression, and Support Vector Classifier for predicting the risk level of heart disease. The system uses 11 medical parameters such as age, gender, blood pressure, cholesterol for prediction. The Heart Disease Prediction System can be considered as a preliminary test for the likelihood of patients being diagnosed with heart disease. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

Keywords: Data Mining, KNN Algorithm, Logistic Regression, Support Vector Classifier, Disease Diagnosis.

Introduction

It's estimated that around 17.5 million deaths occur due to heart attacks each year worldwide. Early prediction & proactive precautions are crucial for heart attacks & other heart-related diseases. Most heart patients share certain common traits. These traits can be used for early prediction. Accurate data can not only avoid the wrong diagnoses but also save human resources. False-positive diagnosis of heart disease will lead to unnecessary panic & vice versa will result in a lack of precautions thereby reducing the chance of survival. A wrong diagnosis is painful to both patients & hospitals. This leads to concerns over health problems which are expected to increase in coming years. With accurate predictions, we can solve unnecessary trouble. If data mining techniques can identify certain common factors such as cholesterol, high blood pressure, etc, heart failure can be predicted & we can bring down its risk. Moreover, if these characteristics are used as preliminary testing; complicated & costly diagnosis processes can be avoided by many; which in turn allows access to serious patients for such treatment. This data set is very limited & the conclusions are weak but the techniques used can be used in a larger data set & allow us to produce more accurate conclusions.

Broad Context

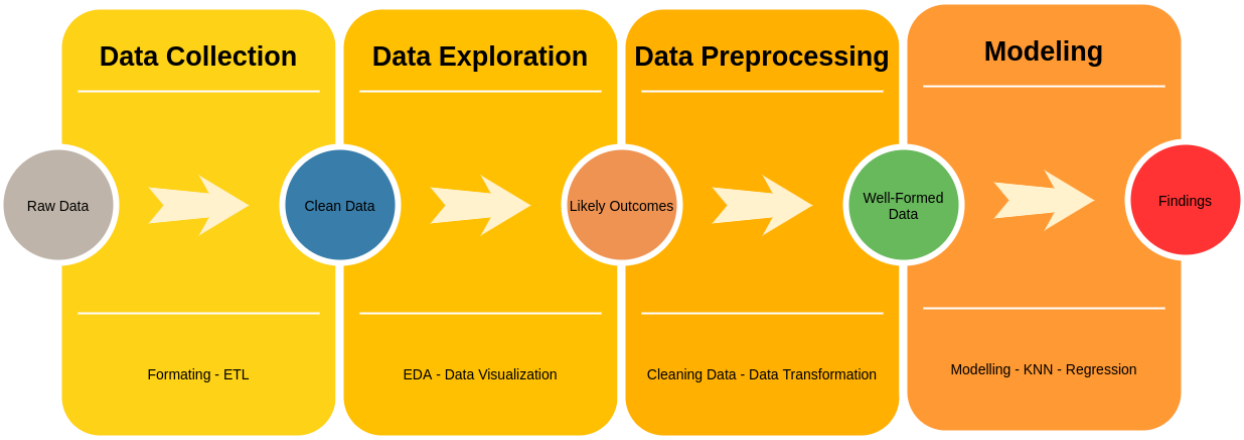
In order to analyze & predict heart failure, algorithms have to consider various factors that affect the heart such as cholesterol, blood pressure, heart rate, etc. Here is the list of various factors which are used as features in the dataset:

Attribute	Description	Values
age	Age of a person (# days)	39 to 65 years
gender	Gender (binary int)	1 - female, 2 - male
height	Height (discrete int, cm)	68cm - 197cm
weight	Weight (discrete int,kg)	28kg - 180kg
ap_hi	Systolic bp (discrete int, mmHg)	11 mmHg - 806 mmHg
ap_lo	Diastolic bp (discrete int, mmHg)	0 mmHg - 8099 mmHg
cholesterol	Cholesterol (discrete int)	1 - normal 2 - above normal 3 - well above normal
gluc	Glucose (discrete int)	1 - normal 2 - above normal 3 - well above normal
smoke	Smoking Habit (binary int)	1 - Yes, 0 - No
alco	Alcohol Intake (binary int)	1 - Yes, 0 - No
active	Physical activity (binary int)	1 - Yes, 0 - No
cardio	Presence or absence of cardiovascular disease (binary int)	1 - Yes, 0 - No

As represented above the input database uses 12 features. Out of these 5 are continuous & 7 discrete. The target feature will be cardio, since it is a binary feature, the models will be using **classification** to draw conclusions.

Study System

The study will be conducted in 3 phases: the pre-processing phase, where we chose the most relevant attributes, the second phase uses simple techniques to narrow down possible outcomes, & the third phase applies Machine Learning algorithms in order to select those attributes that give better accuracy. Our proposal is divided into several phases, the approach is shown below:



Each individual step within the four phases are detailed as follows:

1. **Dataset**

   Dataset can be summarized into 12 columns, containing eleven features and the corresponding **Cardiovascular Disease** possibility. The entire data consists of about 3500 entries containing both male and female genders of different age categories. Each of the features is differentiated into numerical and categorical values accordingly.

```
df.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
27504	57	1	164	65.0	120	80	1	1	0	0	1	1
57935	54	1	165	66.0	12	80	1	1	0	0	1	0
40387	47	1	161	77.0	130	90	1	1	0	0	1	1
24455	62	2	170	78.0	140	90	1	2	1	0	1	1
66362	42	2	178	70.0	120	80	1	1	1	1	1	0

Dataset dimensions consist of 3500 rows and 12 columns. (3500 x 12)

2. **Data Exploration/Summarization**
   Since the dataset consists of a rational amount of samples and features, it is necessary to analyze, identify and infer relationships within the data and draw conclusions from it, before moving to the pre-processing step.

❖ **The statistical description of features with their respective percentiles**

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
age	3500.0	52.642000	6.705166	39.0	48.0	53.0	58.0	64.0
gender	3500.0	1.344000	0.475109	1.0	1.0	1.0	2.0	2.0
height	3500.0	164.511714	8.379887	68.0	160.0	165.0	170.0	197.0
weight	3500.0	74.530800	14.857876	28.0	65.0	72.0	83.0	180.0
ap_hi	3500.0	126.964857	21.519690	11.0	120.0	120.0	140.0	806.0
ap_lo	3500.0	94.150571	168.978761	0.0	80.0	80.0	90.0	8099.0
cholesterol	3500.0	1.360000	0.680097	1.0	1.0	1.0	1.0	3.0
gluc	3500.0	1.240286	0.584872	1.0	1.0	1.0	1.0	3.0
smoke	3500.0	0.090857	0.287447	0.0	0.0	0.0	0.0	1.0
alco	3500.0	0.054571	0.227174	0.0	0.0	0.0	0.0	1.0
active	3500.0	0.791429	0.406345	0.0	1.0	1.0	1.0	1.0
cardio	3500.0	0.486571	0.499891	0.0	0.0	0.0	1.0	1.0

❖ **Number of Unique values belonging to each feature**

```
n = df.nunique(axis=0)
n
```

```
age          26
gender        2
height       59
weight      116
```

The statistical summary indicates that there are about 26 different age categories combining both genders. It can also be inferred from the summary that the dataset consists of both numerical and categorical values.

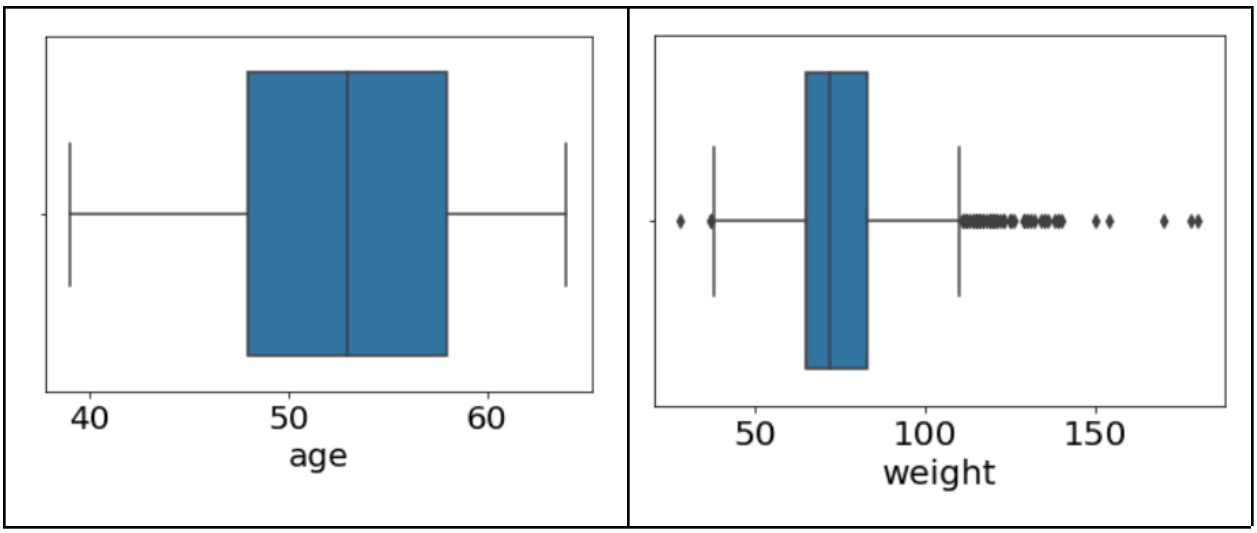
❖ **Gender Distribution**

```
df['gender'].value_counts()
```

```
1    2296
2    1204
Name: gender, dtype: int64
```

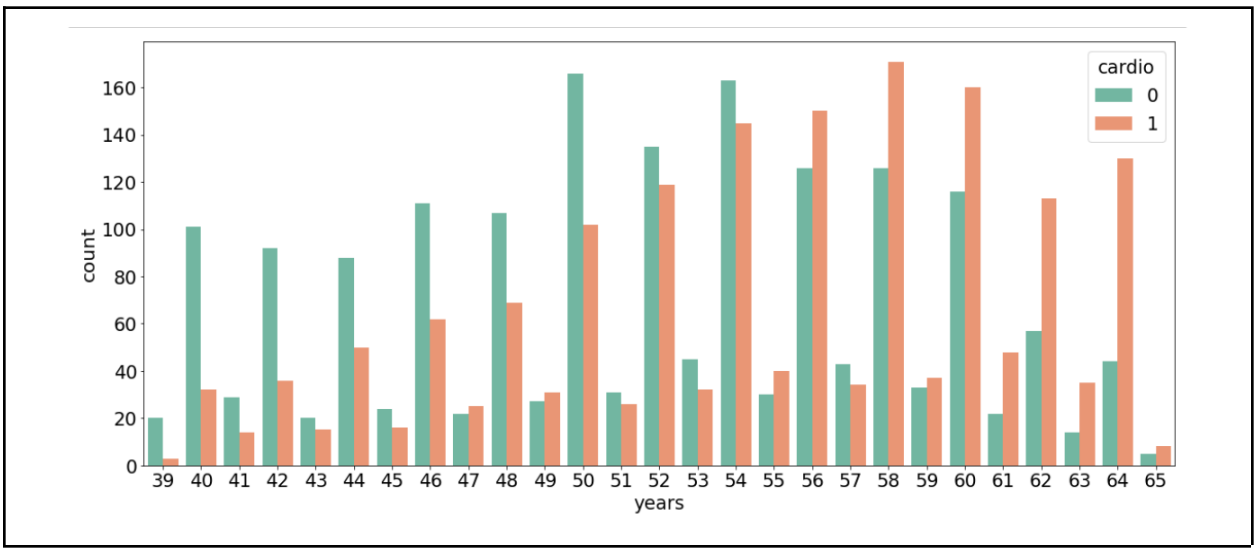
The dataset consists of 3500 samples of which 2296 are female samples and 1204 are male samples. Visual representation of the distribution is shown above where females are represented as 1, and males are represented as 2.

❖ **Boxplot Representation of Age & Weight features**



The above boxplot determines the distribution of the mentioned attributes, age, and weight. Age displays an approximate symmetric plot while the weight is right-skewed with the majority of data distributed between 40 Kg to 110 Kg.

❖ **Cardio Graph over the age (years)**

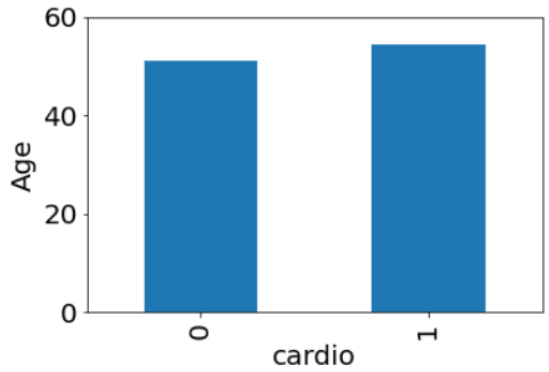


It can be observed that people over the age of 55 are more exposed to CVD.

❖ **Comparing the average age of people with CVD & the average age of healthy people.**

```
df.groupby('cardio')['age'].mean()
```

```
cardio
0    51.087368
1    54.282443
Name: age, dtype: float64
```



It can be observed that the average age of people with Cardiovascular Disease (CVD) is slightly higher than that of healthy people.

❖ **Comparing the number of smokers and non-smokers**

```
df['smoke'].value_counts()
```

```
0    3182
1     318
Name: smoke, dtype: int64
```

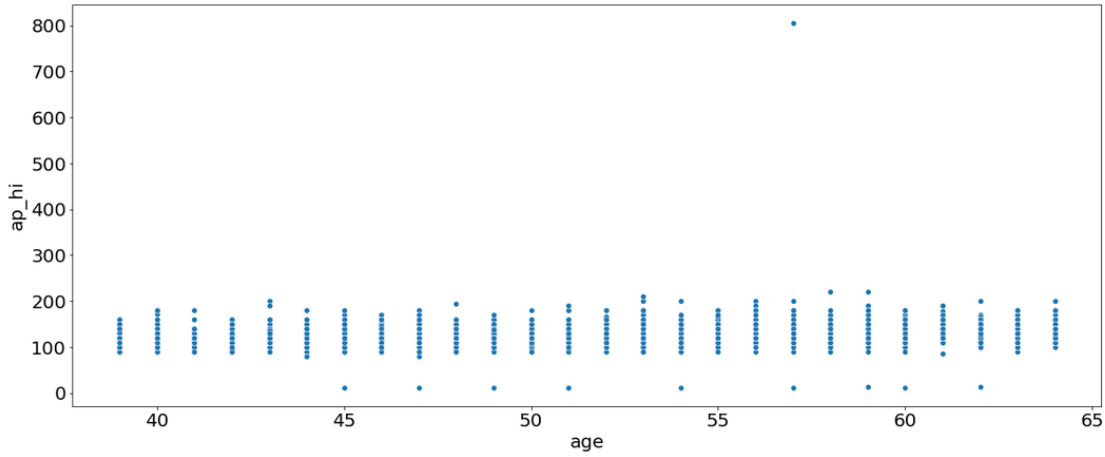
There are about 3182 non-smokers and 318 smokers within the dataset, represented by 0 and 1 respectively.

❖ **The average age of people with CVD who are also smokers**

```
df[(df['smoke'] == 1) & (df['cardio'] == 1)]['age'].mean()
52.64492753623188
```

The average age of people who are both smokers and have CVD is 52-53.

❖ **A scatterplot showing high blood pressure variance with age**



The blood pressure is found high over the age group of 53 - 58.

3. Data Preprocessing

It is an important step in Machine Learning as the quality of the data and the useful information that can be derived from it directly affects the ability of our model to learn. Therefore, it is extremely important that we preprocess our data before feeding it into our model.

The initial dataset chosen has 70000 rows of data, the high count would make it difficult to process. Thus we use the *train_test_split* method to split the data into 3500 rows. The *id* feature was removed from the dataset. The *age* attribute which was initially given in - number of days was converted to number of years, after dividing the days by 365 and rounding off the values.

The dataset is then split into X & Y lists, where X contains all (but *cardio*) features and Y contains the feature *cardio*. X is standardized. Both the lists are split to train and test groups by the *train_test_split* method.

4. **Modeling**

This is the phase where we apply and test the chosen algorithms (KNN, Logistic Regression, SVM) to find the best between them. The main goal is to find an algorithm that remains stable while training and testing. We finish our proposal by selecting the best algorithm that gives the best accuracy.

```
models = {
    "SVM": lambda x, y: svm.SVC().fit(x, y),
    "Logistic Regression": lambda x, y: linear_model.LogisticRegression().fit(x, y),
    "KNN": lambda x, y: neighbors.KNeighborsClassifier(n_neighbors=5).fit(x, y)
}
```

Methods

KNN Algorithm

The KNN algorithm finds the K closest neighbors of a given test data point by calculating the **Euclidean Distance Measure** with all the other training data samples. In the case of classification, the majority class of the labels containing the indices of K closest distance samples is used to predict Cardiovascular Disease(CVD) possibility.

Logistic Regression

Logistic Regression is a classification algorithm, which uses the concept of probability. It uses a complex cost function defined as a **Sigmoid** or a **Logistic** function that **classifies** the data. The limits of the cost function always range from 0 to 1.

The aim is to optimize the cost function such that all the data points converge to their entitled class; this is done by the **gradient descent**

algorithm. The prediction as mentioned above can be classified, based on the threshold value, which decides whether the person is likely to fall into a range where the chances of heart failure are possible.

Support Vector Classifier:

Support Vector Machines are supervised learning methods used for classification, regression as well as outlier detection. Support Vector Classifier tries to find the optimal hyperplane that can classify the instances with the least error. It tries to fit the best line within a threshold value, that is; the distance between the hyperplane and boundary line. The algorithm optimizes the distance between the nearest classes. This is done by finding the maximum distance with the Convex Optimisation problem.

K-fold Cross-Validation

K-fold cross-validation is one of the most commonly used model evaluation methods. Even though this is not as popular as the validation set approach, it can give us a better insight into our data and model.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. Every split of the data forms folds, containing a proportion of training and test data. Each fold is trained and tested using the models and the best one is chosen.

We have applied k-fold cross-validation for all three models of our data. Implementation of K-fold Cross Validation is given below:

```
def kfold(x, y, func, val = 5):
    models = []
    kf = model_selection.KFold(n_splits=val)
    for idx in kf.split(x):
        KNN = func(x[idx[0]], y[idx[0]])
        models += [[KNN, metrics.accuracy_score(y[idx[1]], KNN.predict(x[idx[1]]))]]
    return sorted(models, key=lambda a: a[1])[-1]
```

Results

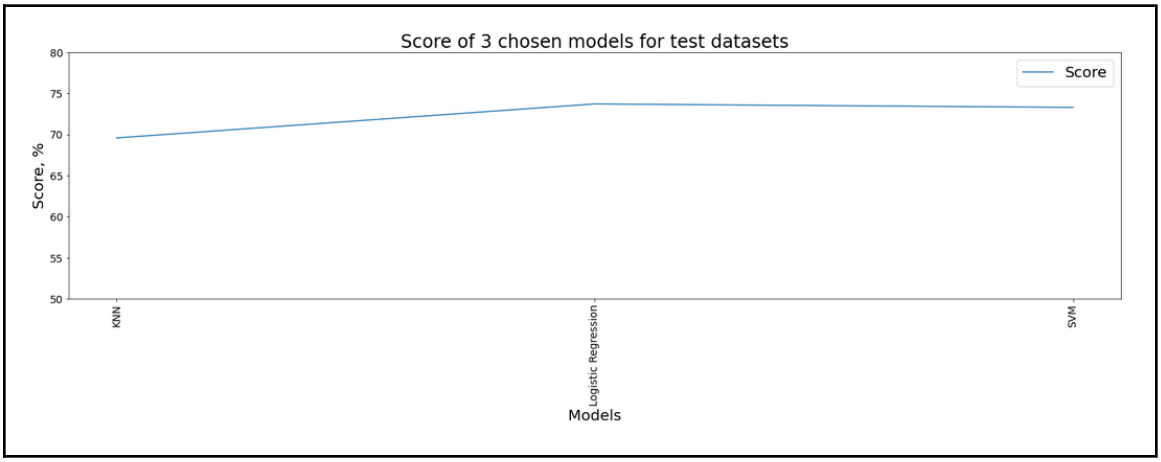
After applying KNN, Logistic regression, and Support Vector Machine algorithm on the training dataset, we tested the accuracy of each algorithm on the test dataset using the accuracy_score method. The below table displays the accuracy scores of each algorithm :

	KNN	LR	SVM
Accuracy	69.57%	73.71%	73.29%

```
model_scores, model = [], []
for name, fn in models.items():
    best_model, best_out = kfold(X_train, Y_train, fn)
    Y_pred = best_model.predict(X_test)
    model_scores += [to_per(metrics.accuracy_score(Y_test, Y_pred))]
    model += [best_model]
    print(f"score: \t{model_scores[-1]}% {name}")
```

```
score:  69.57% KNN
score:  73.71% Logistic Regression
score:  73.29% SVM
```

After analyzing the above accuracy scores, it has been found that Logistic regression is the best algorithm. It resulted in the highest accuracy among the 3 algorithms with a score of 72%. A graph comparing the accuracy scores of different models is shown below.



Discussion

Cardiovascular diseases have been one of the most affecting diseases around the globe. Therefore, it has become important to develop high accuracy models which can predict a person's chance of heart failure well in advance. Our dataset gives a rough view of the possibility of cardiovascular diseases among people. Throughout our discussions, we came to see how different Machine Learning models were able to give around about 65-73 percent accuracy for our dataset.

Conclusion

Heart disease is one of the most frequently detected diseases. Therefore, it has become important to develop high accuracy models which can predict a person's chance of heart failure well in advance. In this study, we have used a dataset to analyze the correlation between the features and target label(i.e predicting heart failure). We have visualized these using boxplot, line graph, bar graph,..etc.

Finally, we chose 3 Machine Learning algorithms and concluded with the best predicting model. **We were able to infer that features such as cholesterol, high blood pressure, age, etc were the dominant ones leading to cardiovascular diseases.**

feature	weight
age	0.371671
gender	0.013468
height	0.010407
weight	0.236558
ap_hi	1.162477
ap_lo	-0.043142
cholesterol	0.407737
gluc	-0.155259
smoke	-0.139625
alco	0.022211
active	-0.018056