Heart Disease Classifier

Ayush Arora, Muriel Banze, Shweta Wahane

Agenda

1. Motivation
2. Problem definition
3. Key issues and Alternative ways
4. Related work and Limitations
5. Our Approach
6. Validation
7. Conclusion
8. Future work
9. References

1. Motivation

- Heart is one of the most important organ of the body.
- Around 17 million people die due to cardiovascular diseases every year.
- It is equivalent to about 29.2% deaths each year.
- Heart Ailments are result of unhealthy lifestyles that lead to high blood pressure, diabetes, cholesterol fluctuation, exhaustion etc.

2. Problem Definition

- An early diagnosis of Heart related ailments is necessary for urgent health care.
- Running all the tests physically on every patient will be a cost inefficient and troublesome way of providing the care.
- The probability of human error exists even under highly trained professionals.
- A number of tests can then be performed on the potential patients from time to time to reduce the effect of the disease.
- Use of data analytics tools applied to the available data can help health care providers identify the early signs of Heart Ailments.

3. Key Issues and alternative ways

- For all the existing systems in the medical domain the biggest issue is the availability of data in the appropriate amount to train the predictive model.
- Also there is no standardization in the process of quantification and storage of the data which makes the pre processing of data necessary and troublesome.
- Doctors have trouble keeping up with the upcoming technologies.
- The best possible solution to this problem is to have a standard to store this data from as many hospitals.
- This way seems to be naive and laborious.
- Hence we provide a way in the project which could help the modelling and prediction of Medical data a little better.

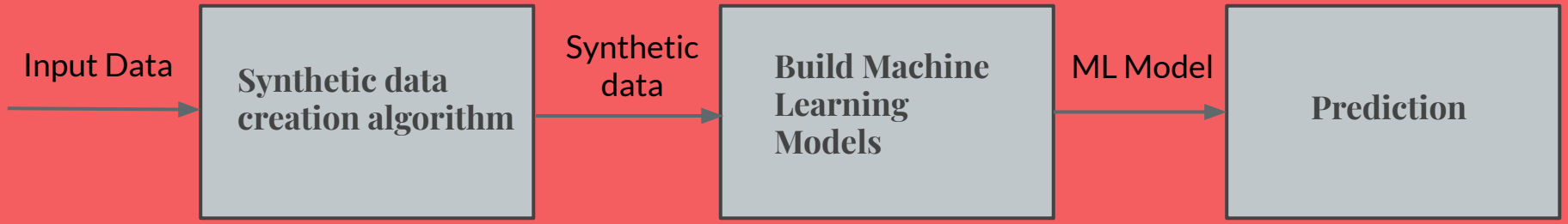
4. Related Work & Limitations

- R. Kavitha et al [3], uses a larger source of raw data to build the models and uses Principal Component Analysis for dimensionality reduction. The models build with this methodology have achieved good accuracies but at the expense of interpretability of the model.

- K. Srinivas et al [4], contrasts various algorithms run on a clinical data set with an aim to determine the probability of a heart problem based on feature extraction and pattern generation. However, the focus is not on improving the efficiency of the algorithms and the interpretability of the model.

5. Our Approach

5.1 Proposed System



5.2 Input Data

- The dataset used is the Cleveland Heart Disease Dataset from UCI Repository.
- It contains 303 data instances, each of which is a patient record.
- There are 14 attributes for both male and female patients with readings like blood pressure, cholesterol level, blood sugar level, maximum heart rate, number of major vessels, etc.

5.3 Creation of Synthetic Data

- The synthetic dataset is generated using the Synthpop library in R.
- It creates a synthetic version of the dataset of same size while retaining the same properties and relationship between the variables.
- Finally a final dataset is created by merging the rows in the original and the synthetic dataset.

	Age	Sex	Chest_Pain	Resting_Blood_Pressure	Cholestrol_Level	Fasting_Sugar	Restinc
1	67	1	4	160	286	0	
2	67	1	4	120	229	0	
3	37	1	3	130	250	0	
4	41	0	2	130	204	0	
5	56	1	2	120	236	0	
6	62	0	4	140	268	0	

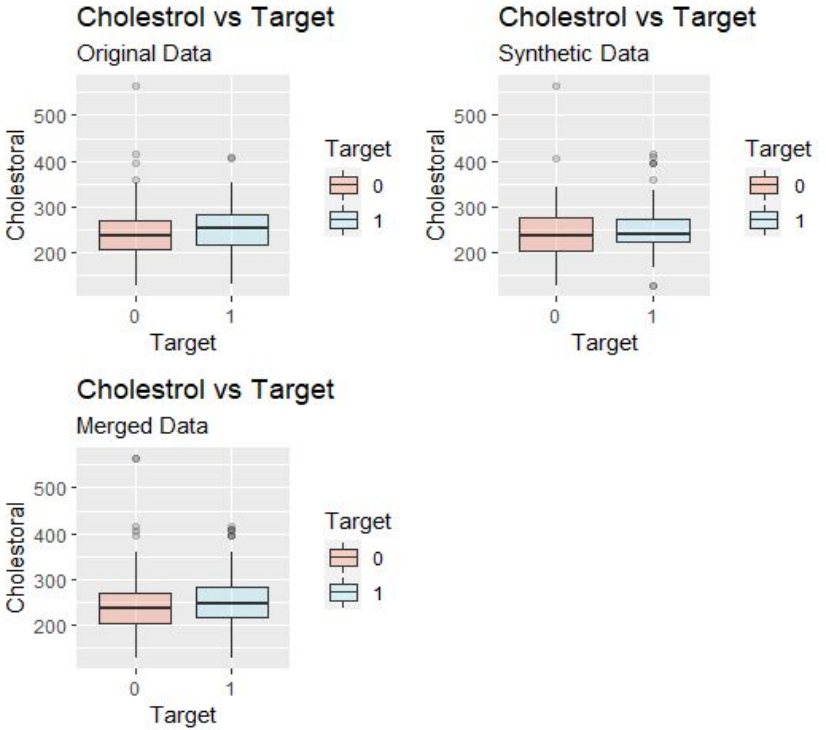
	Max_Heart_Rate	Exercise_Induced_Anigna	ST_Depressions	ST_Depressions_Slope
1	108	1	1.5	2
2	129	1	2.6	2
3	187	0	3.5	3
4	172	0	1.4	1
5	178	0	0.8	1

	Age	Sex	Chest_Pain	Resting_Blood_Pressure	Cholestrol_Level	Fasting_Sugar	Resting_ECG
1	51	1	4	130	274	0	2
2	48	1	4	112	204	0	2
3	57	1	2	124	284	0	0
4	62	1	3	140	304	0	2
5	66	1	4	120	229	1	2
6	50	0	3	105	340	1	0

	Max_Heart_Rate	Exercise_Induced_Anigna	ST_Depressions	ST_Depressions_Slope
1	168	0	0.5	1
2	160	0	0.6	2
3	147	0	2.6	1
4	145	0	1.8	2
5	132	1	1.2	1
6	132	0	1.8	2

(Fig 1) head(Original_data) versus head(synthetic_data) created.

5.4 Data Visualization



(Fig 2) Snapshot displaying Cholesterol vs Target

(Fig 3) Snapshot displaying Sugar vs Target

5.5 Performance Measures

- We use Test Prediction Accuracy and Error Rate as the measures.

- Accuracy is defined as the percentage of correct predictions for the **test** data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

- Error Rate is the percentage of incorrect predictions for the **test** data. It is calculated by dividing the number of incorrect predictions by the number of total predictions.

5.6 Algorithms Executed

- k-Nearest Neighbours
- Support Vector Machines
- Logistic Regression
- Quadratic Discriminant Analysis
- **Decision Tree**

5.6.1 k-Nearest Neighbors

- KNN works on a principle assuming every data point falling near to each other is falling in the same class. That means similar things are near to each other.

- KNN algorithm is initialized with a k value which is the nearest Neighbor to that data point which is to be classified. We set the value of k as 6 it will looks for 6 nearest Neighbors to a particular data point.

- **Results:**

```
> print(table(knn_model_3, merged_test$Target))

knn_model_3  0  1
          0 59 22
          1 25 42

> print(paste("Accuracy of kNN on Merged Dataset = ", (59 + 42)*100/148,"%"))
[1] "Accuracy of kNN on Merged Dataset =  68.2432432432432 %"
> print(paste("Error Rate of kNN on Merged Dataset = ", (22 + 25)*100/148,"%"))
[1] "Error Rate of kNN on Merged Dataset =  31.7567567567568 %"
```

(Fig 4) Snapshot displaying the confusion matrix for the KNN merged dataset

5.6.2 Support Vector Machines

- SVM is an algorithm that takes the data as an input and outputs a line or a hyperplane that splits the data into separate classes.

- **Results:**

```
> print(table(merged_test[,14], pred ))
   pred
     0  1
  0 73 11
  1 12 52
```

```
> print(paste("Accuracy of SVM on Merged Dataset = ", (73 + 52)*100/148,"%"))
[1] "Accuracy of SVM on Merged Dataset =  84.4594594594595 %"
> print(paste("Error Rate of kNN on Merged Dataset = ", (11 + 12)*100/148,"%"))
[1] "Error Rate of kNN on Merged Dataset =  15.5405405405405 %"
```

(Fig 5) Snapshot displaying the confusion matrix for the Support Vector Machine merged dataset

5.6.3 Logistic Regression

- Logistic Regression uses a logistic function for predicting a target variable. The response variable that is binary belongs either to one of the classes. This algorithm computes probability values that range from 0 and 1.

- **Results:**

```
> print(table(merged_test$Target, predict > 0.7))

    FALSE TRUE
  0    79    5
  1    21   43


> print(paste("Accuracy of LogR on Merged Dataset = ", (79 + 43)*100/148,"%"))
[1] "Accuracy of LogR on Merged Dataset =  82.4324324324324 %"
> print(paste("Error Rate of LogR on Merged Dataset = ", (21 + 5)*100/148,"%"))
[1] "Error Rate of LogR on Merged Dataset =  17.5675675675676 %"
```

(Fig 6) Snapshot displaying the confusion matrix for the Logistic Regression merged dataset

5.6.4 Quadratic Discriminant Analysis

- QDA algorithm is used to find a quadratic combination of features that characterizes or separates two or more classes of objects or events.

- **Results:**

```
> print(table(qda_merged_prediction$class, merged_test$Target))

    0  1
  0 73 10
  1 11 54

> print(paste("Accuracy of QDA on Synthetic Dataset = ", (73 + 54)*100/148,"%"))
[1] "Accuracy of QDA on Synthetic Dataset =  85.8108108108108 %"
> print(paste("Error Rate of QDA on Synthetic Dataset = ", (10 + 11)*100/148,"%"))
[1] "Error Rate of QDA on Synthetic Dataset =  14.1891891891892 %"
```

(Fig 7) Snapshot displaying the confusion matrix for the Quadratic Discriminant Analysis merged dataset

5.6.5 Decision Trees (Core Algorithm)

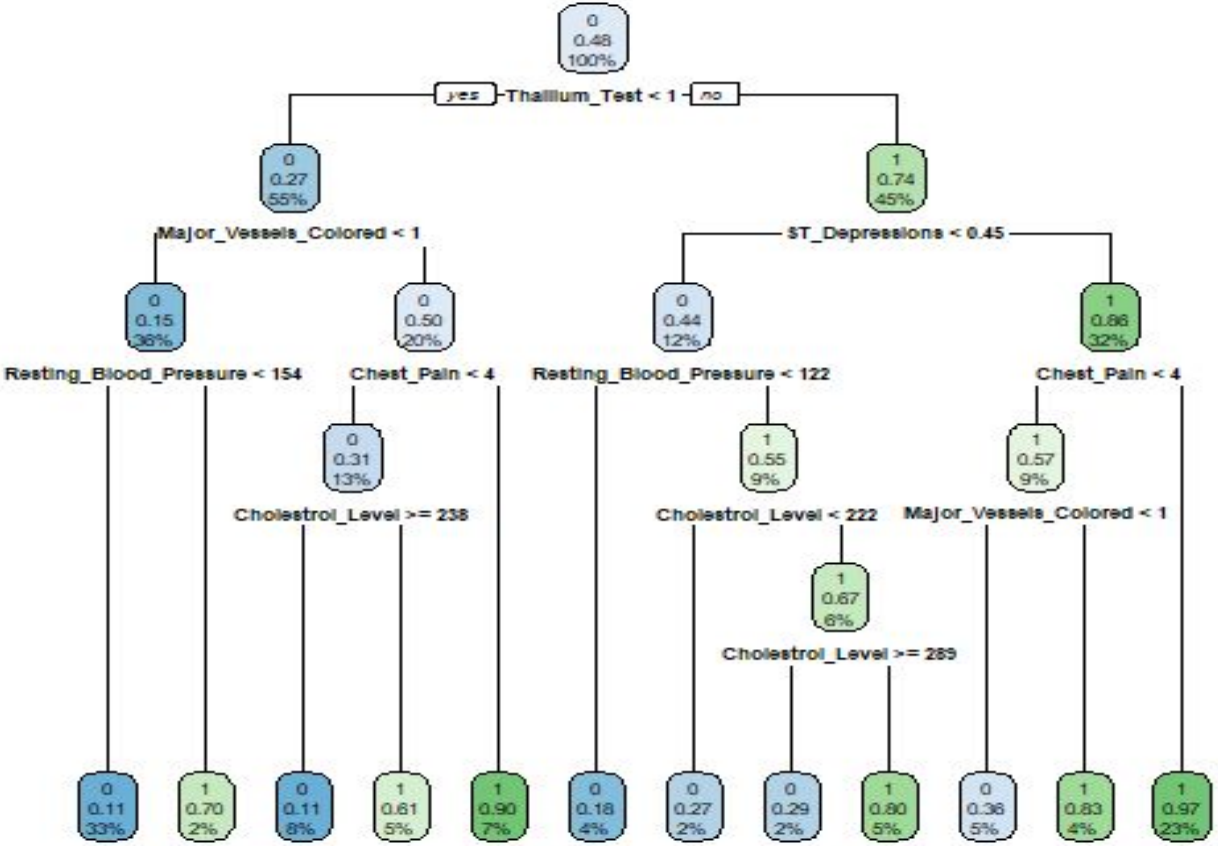
- Decision tree algorithm generates a graph to represent choices and their results in form of a tree.
- The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.
- The output of the decision tree is very simple to understand and interpret for the human beings.
- Doctors while using a decision making system for medical decisions prefer to know how the model derives to the result instead of just receiving outputs from a black box.
- **Results:**

```
> print(table(merged_predictions, merged_test$Target))

merged_predictions  0  1
                  0 73 11
                  1 11 53

> print(paste("Accuracy of Decision Tree on Merged Dataset = ", (73 + 53)*100/148,"%"))
[1] "Accuracy of Decision Tree on Merged Dataset =  85.1351351351351 %"
> print(paste("Error Rate of Decision Tree on Merged Dataset = ", (11 + 11)*100/148,"%"))
[1] "Error Rate of Decision Tree on Merged Dataset =  14.8648648648649 %"
```

(Fig 8) Snapshot displaying the confusion matrix for the Decision Trees merged dataset

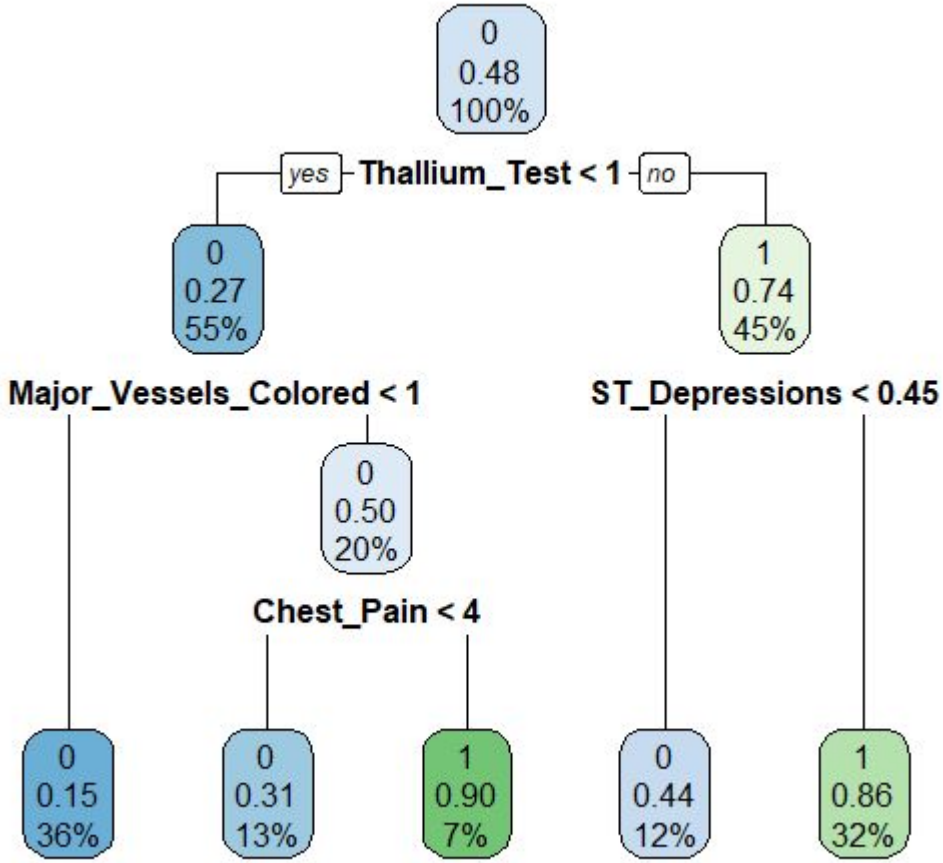


(Fig 9) Snapshot displaying the Decision Tree

5.6.6 Decision Trees Fine Tuning

Pruning:

- The tree algorithm repeatedly partitions data into smaller and smaller subsets until those final subsets are homogeneous in terms of the outcome variable.
- This often leads to the final leaves consisting of only one or a few data points which signifies overfitting in the model.
- To overcome this we prune the decision tree to reduce the size of the trees by removing sections of the tree that are non-critical and redundant to classify instances.



(Fig 10) Snapshot displaying the results after pruning the decision tree

5.6.6 Decision Trees Fine Tuning

Bagging:

- Bagging or Bootstrap Aggregation works on the principles of bootstrapping which creates several subsets of the training data sample chosen randomly with replacement.
- All the data sample are used to train a large number of decision trees. This results in ensemble of different models.
- This reduces the variance of the decision tree as the data doesn't overfit anymore.

5.7 Observations

Algorithm	Accuracy/ Error Rate Original	Accuracy/ Error Rate Synthetic	Accuracy/ Error Rate Merged
Decision Trees	72.97/ 27.02	78.37/ 21.62	85.13/ 14.86
QDA	77.02/ 22.97	77.02/ 22.97	85.81/ 14.18
k-Nearest Neighbours	63.51/ 36.48	63.51/ 36.48	68.24/ 31.75
Support Vector Machines	82.43/ 17.56	79.72/ 20.27	84.59/ 15.54
Logistic Regression	74.32/ 25.67	74.32/ 25.67	82.43/ 17.56

6. Validation

Fine-tuned Decision Tree Algorithm	Accuracy/ Error Rate Merged	Accuracy/ Error Rate Original
87.83/ 12.16	85.13/ 14.86*	72.97/ 27.02

The final pruned and bagged decision tree with the merged data shows an improved accuracy of 87.83% for the merged test dataset which is about 2.7% better than the initial model.

7. Conclusions

- We have examined that Thallium Test, Major Vessels Colored, Chest Pain, and ST Depressions at a particular threshold are the significant predictors.

- A patient with Chest Pain greater than 4 units will suffer from a Heart-related issue with a probability of 0.90 and one with ST Depressions value above 0.45, has a probability of 0.86 of suffering from a Heart Disease.

- Training the model with synthetic datasets and fine tuning the decision trees algorithm increases the accuracy of the system distinctly and provides a good level of interpretability.

8. Future Work

We aim to integrate the models built with a user interface.

This would help in predicting heart ailments for individual patients.

We also aim to collect and also synthesize more data to train the models more efficiently.

9. References

[1] https://archive.ics.uci.edu/ml/datasets/heart+disease

[2] https://www.medicalnewstoday.com/articles/237191#news

[3] Srinivas, K. & Rani, B. & Govardhan, Dr. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering. 2. 250-255.

[4] R. Kavitha and E. Kannan, 'An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining', in *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, India, 2016, pp. 1–5.

[5] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

[6] Nowok B., Raab G. M, Dibben C., "synthpop: Bespoke Creation of Synthetic Data in R"

[7] Sabay, Alfeo; Harris, Laurie; Bejugama, Vivek; and Jaceldo-Siegl, Karen (2018) "Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data," SMU Data Science Review: Vol. 1 : No. 3 , Article 12.

Thank You!