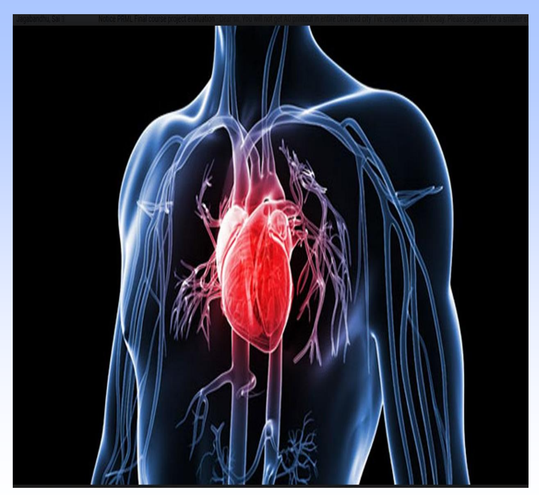
Heart Disease Classification

Lokesh Kumar Nirania
Karthik Vadlamudi
Kunal Kumar
Unnati Athwani
Indian Institute of Technology Dharwad
Computer Science & Engineering
Email: 1700100-09/16/12/06@iitdh.ac.in



Abstract

We used SVM Algorithm with regularization parameter 'lambda' to classify patients as 'may have heart attack' and 'may not have an heart disease'

We also used various kernels like 'liblinear', 'sag' in Logistic Regression, 'sigmoid', 'rbf', 'poly', 'linear' and compared accuracy by splitting data randomly in a **80:20** ratio of **train : test** proportion

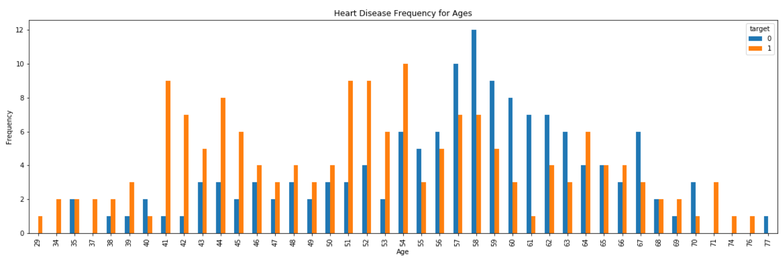
Before training data, we normalized it using 'min-max' and 'z-score' methods

Introduction

Of all the applications of machine-learning, diagnosing any serious disease using a black box is always going to be a hard shell. If the output from a model is the particular course of treatment (potentially with side-effects), or surgery, or the *absence* of treatment, people are going to want to know **why**.

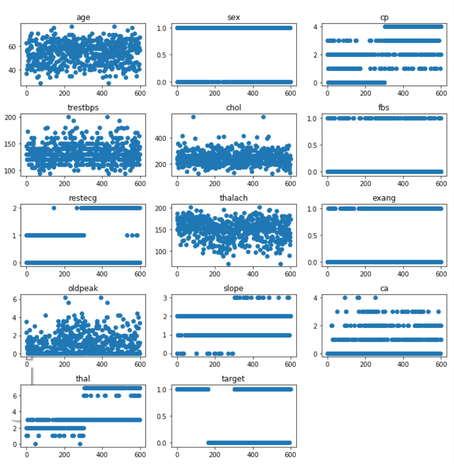
The dataset we used have a number of variables along with a **target** condition of **having** or **not having heart disease**

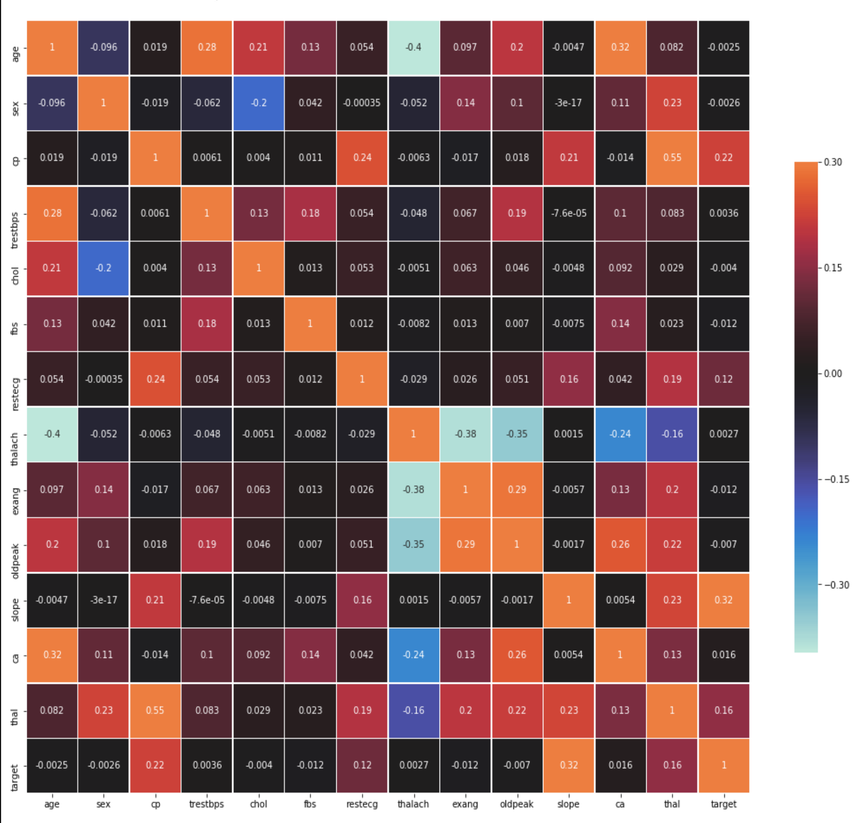
Below is a graph plotting the frequencies at which people of a certain age have been and have not been diagnosed with heart diseases according to a research data



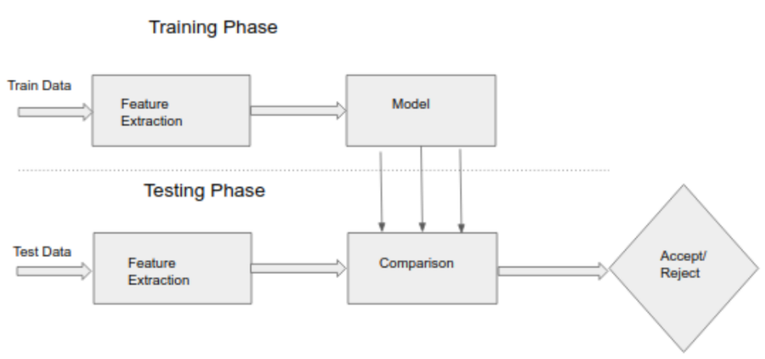
Data Preprocessing

- We imported the data from raw csv file to python list
- We **normalized** our data using **'z-score'** and **'min-max'** method
- We wanted to **reduce the features**
- Right hand side is **visualization of spread of each entry** (5 different types of entry 'cp' is in our data)
- **'target'** contains the info whether the patient will have heart attack or not
- We performed **feature extraction** in which we keep only useful features by constructing **covariance matrix** which **numerically describes relation of each entry with every other entry** ('target' in one of the entries)
- We picked only those entries as our features which have **co-relation values** with **3 non-zero significant digits** i.e., **(**value 0.016 is accepted but 0.0025 is discarded as feature**)**





Basic Block Diagram of SV System



Proposed Modeling Scheme

Training Phase :-

These are the set of features we used in our initial data:
age, sex (1 = Male, 0 = Female), **cp:** The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic), **trestbps:**The person's resting blood pressure (mm Hg on admission to the hospital), **chol:** cholesterol in mg/dl, **fbs:** fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false), **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria), **thalach:** maximum heart rate achieved, **exang:** Exercise induced angina (1 = yes; 0 = no), **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot), **slope:** the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping), **ca:** The number of major vessels (0-3), **thal:** A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect), **target:** Heart disease (0 = no, 1 = yes)

However, this, after constructing the co-variance matrix and looking for significant co-relation values, was reduced to the following set features alone:
cp, fbs, restecg, exang, slope, ca, thal and **target**

We have tried running various Logistic Regression Algorithms and various SVM Algorithms as models to train and test the data

Based on the performance, we have chosen SVM algorithm with Gaussian Kernel as the final model to train and test the data as it gives the best performance

The similarity function for Gaussian Kernel, f_1 for landmark l^1 is as follows:

- $f_1 = exp(-\frac{||x-l^{(1)}||^2}{2\sigma^2}) = exp(-\frac{\sum_{j=1}^{N}||x_j-l_j^{(1)}||^2}{2\sigma^2})$
- If x very close to l^1, then $f_1 \approx 1$
- If x is far from l^1, then $f_1 = 0$

And the cost function for SVM becomes:

$$\min_w D\sum_{i=1}^{M} y^{(i)} cost_1(w^T f_i) + (1-y^{(i)}) cost_0(w^T f_i) + \frac{1}{2}\sum_{j=1}^{M} w_j^2$$

$$\text{cost}_1(w^T x) = log(\frac{1}{1+e^{-w^T x}})$$

$$\text{cost}_0(w^T x) = log(1-\frac{1}{1+e^{-w^T x}})$$

Now optimization function is $w^T f \geq 0$ for $y = 1$ and $w^T f \leq 0$ for $y = 0$

Our target is to reduce the cost based on this cost function and we obtain at our w vector

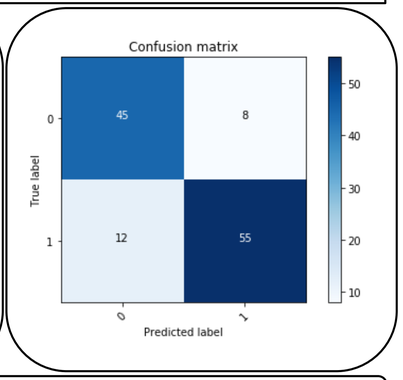
Testing Phase:-

After we have obtained at w vector, we now do our testing part on the 20% test data

For each test datum from this test data, we compute $W^T f$

And if it turns out to be greater than or equal to '0', the output will be '1'(which means, the patient has heart disease)

Otherwise if it turns out to be less than '0', the output will be '0'(which means, the patient does not have heart disease)

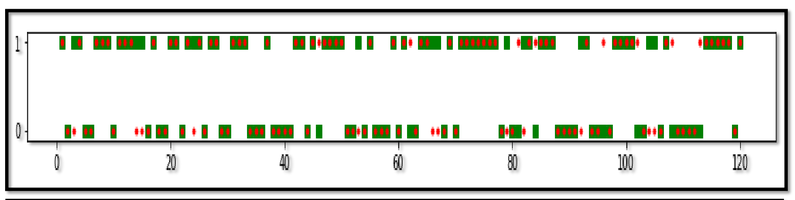


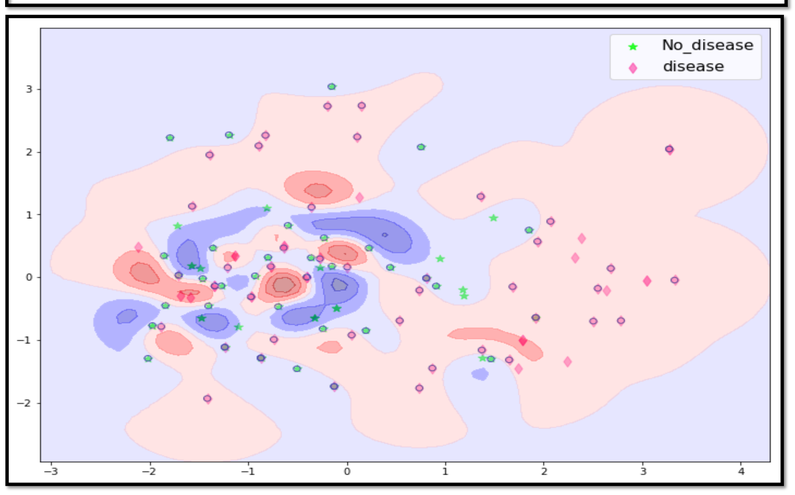
Results

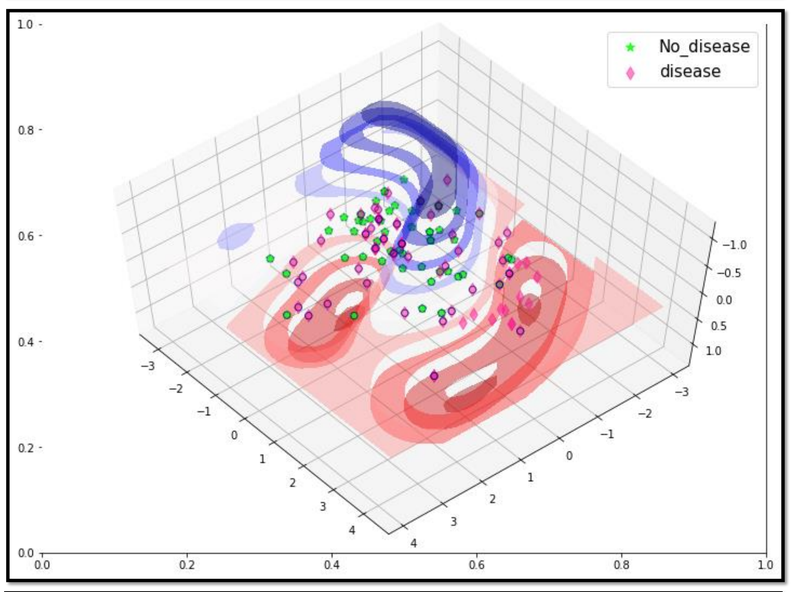
Algorithm	Accuracy		
	Normalisation		Without
	Max-Min	z-score	Normali-
Logistic Regression (liblinear)	66.6	62.5	65
Logistic Regression (sag)	66.7	66	65.1
SVM (rbf)	82.5	83.4	82
SVM (poly)	69.2	75	78.4
SVM (linear)	65.84	65.83	66
SVM (sigmoid)	62.5	65.83	60

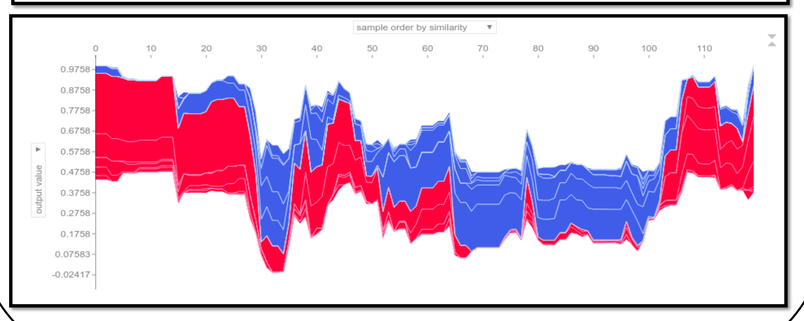
Visualization

Observations are being depicted below :-









Conclusion

We conclude that **cp, fbs, restecg, exang, slope, ca** and **thal** are the main features that determine the chance of being diagnosed with heart diseases.

We can also see that as the number of major blood vessels increases, the probability of heart disease decreases. That makes sense, as it means more blood can get to the heart.

Also, the higher the age, the lower is the chance of heart disease.

Our model can be used for diagnosing heart diseases so that it will be helpful for both health-care professionals and the patients

Acknowledgement

We would like to acknowledge Prof. Sudhanshu Shukla for mentoring us throughout this project. We would also like to thank Prof. SR Mahadeva Prasanna for providing us with this opportunity.

References

[1] We obtained the data used for our project from here:
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] J. Soni et al., "Intelligent and effective heart disease prediction system using weighted associative classifiers," International Journal on Computer Science and Engineering, vol. 3, no. 6, pp. 2385–2392, 2011.

[3] N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbour classification technique," in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26. https://doi.org/10.1145/3175684.3175703.

[4] H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," International Journal of Computer Applications, vol. 168, no. 3, pp. 12–17, Jun 2017.

[5] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, pp. 1–16, 2017. https://doi.org/10.4236/jilsa.2017.91001.