**Project Includes:**

1. **Dataset Details**
2. **Error Rates of Algorithms – Impact of sample size, number of layers and number of nodes, activation functions**
3. **Performance of Algorithms – Confusion matrix, RMSE**
4. **Comparison of Algorithms – Which algorithm works better for which dataset**

**Datasets Used:**

**Heart Disease data** : <http://archive.ics.uci.edu/ml/datasets/heart+Disease>

The database has multiple datasets but the Cleveland processed data is suggested for ML projects. The dataset contains 14 attributes In particular.

The last field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments are concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

14 attributes are:

Only 14 attributes used:

1. #3 (age) - age in years

2. #4 (sex) - sex (1 = male; 0 = female)

3. #9 (cp) - chest pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic

4. #10 (trestbps) - resting blood pressure (in mm Hg on admission to the hospital)

5. #12 (chol) - serum cholestoral in mg/dl

6. #16 (fbs) - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. #19 (restecg) - esting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. #32 (thalach) - maximum heart rate achieved

9. #38 (exang) - exercise induced angina (1 = yes; 0 = no)

10. #40 (oldpeak) - ST depression induced by exercise relative to rest

11. #41 (slope) - the slope of the peak exercise ST segment

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

12. #44 (ca) - number of major vessels (0-3) colored by flourosopy

13. #51 (thal) - 3 = normal; 6 = fixed defect; 7 = reversable defect

14. #58 (num) (the predicted attribute) - num: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

(in any major vessel: attributes 59 through 68 are vessels)

<https://archive.ics.uci.edu/ml/datasets/iris>

Initially the data was available in a comma separated html file – Copy pasted values in excel and delimit separated by ‘commas’, the data was obtained for further use.

**Human Sensing Data:**

The dataset consists of the following variables:

Date Time year-month-day hour: minute: second

Temperature, in Celsius

Relative Humidity, %

Light, in Lux

CO2, in ppm

Humidity Ratio, Derived quantity from temperature and relative humidity, in kg water-vapor/kg-air

Occupancy (Target Variable), 0 for not occupied, 1 for occupied status

**Neural Networks:**

Artificial neural networks (ANNs), a form of connectionism,are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve performance) to do tasks by considering examples, generally without task-specific programming.

**ANN:**

Package used: NeuralNet

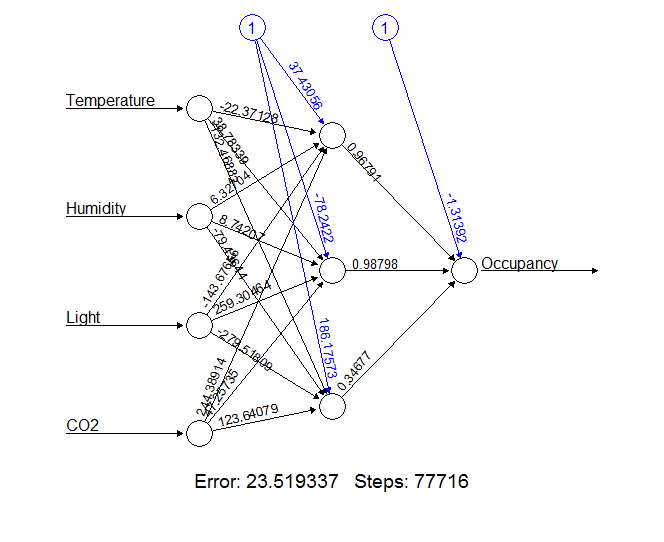
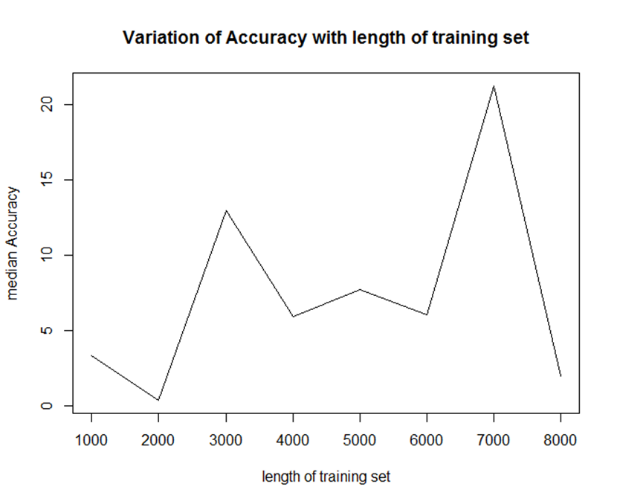
Parameters used:

1. Hidden: To vary layers in neural networks
2. act.fct: Differentiable function used for smoothing the result of the cross product of the covariate or neurons & the weights
3. data: To vary train dataset size

**ANN on Human Sensing Data:**

**Neural Net using 3 variables, 3 hidden layers and logistic function**

Error rate = 23.519; Accuracy is 99%

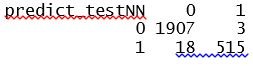
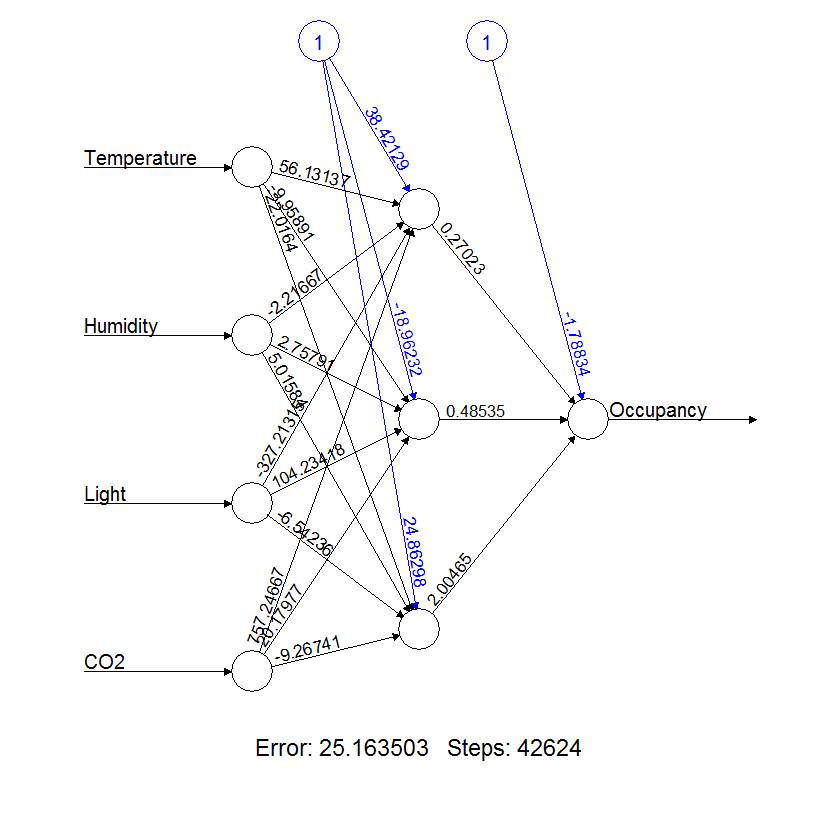
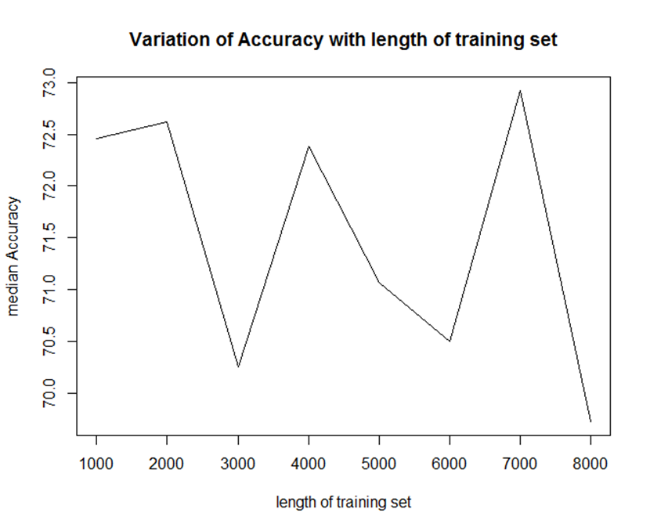


Using logistic function with 3 layers, accuracy that is calculated using the confusion matrix, is maximum when training datasize is 3000 or 7000 – For comparing length of training set with accuracy, k-fold mthodology was used with k=100

**Neural Net using 3 variables, 3 hidden layers and tanh\* function**

\*tanh: tanh is a rescaled logistic sigmoid function

Error rate = 25.16 ; Accuracy is 99.14%

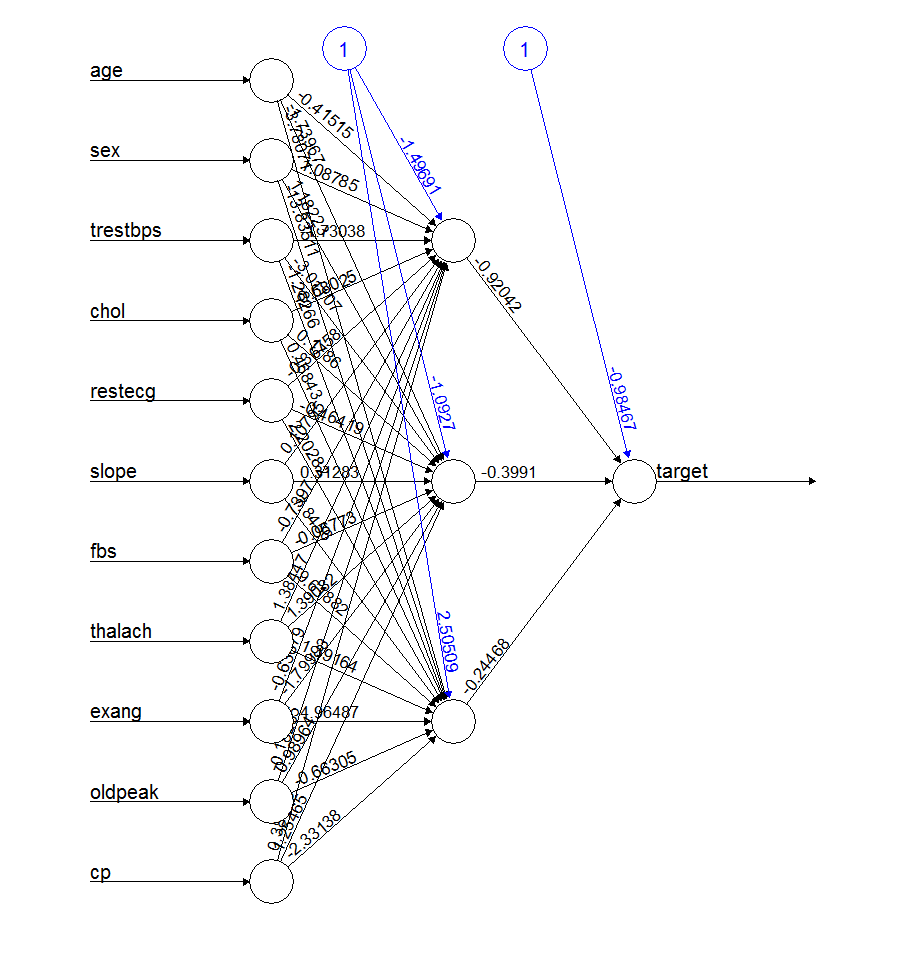
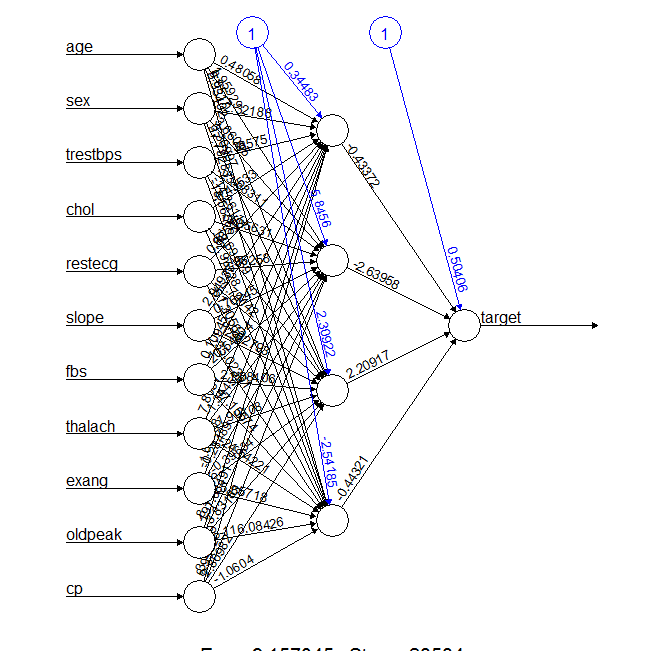


Using tanh function with 3 layers, accuracy that is calculated using the confusion matrix, is maximum when training datasize is 4000 or 7000 – For comparing length of training set with accuracy, k-fold mthodology was used with k=100

**ANN on Heart Disease Data:**

**Neural Net using 3 variables, 3 hidden layers and logistic and tanh function**

Error rate= 51.33 Accuracy=71.2% Error rate=43.79 Accuracy=79.83%



**ANN works better on the Human Sensing dataset**

**k-nearest neighbors:**

In pattern recognition, the k-nearest neighbors’ algorithm (k-NN) is a non-parametric method used for classification and regression. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

**KNN:**

Package used: kknn

Parameters used:

1. k: Number of clusters
2. train: To vary train dataset size

**KNN on Heart Disease Dataset**

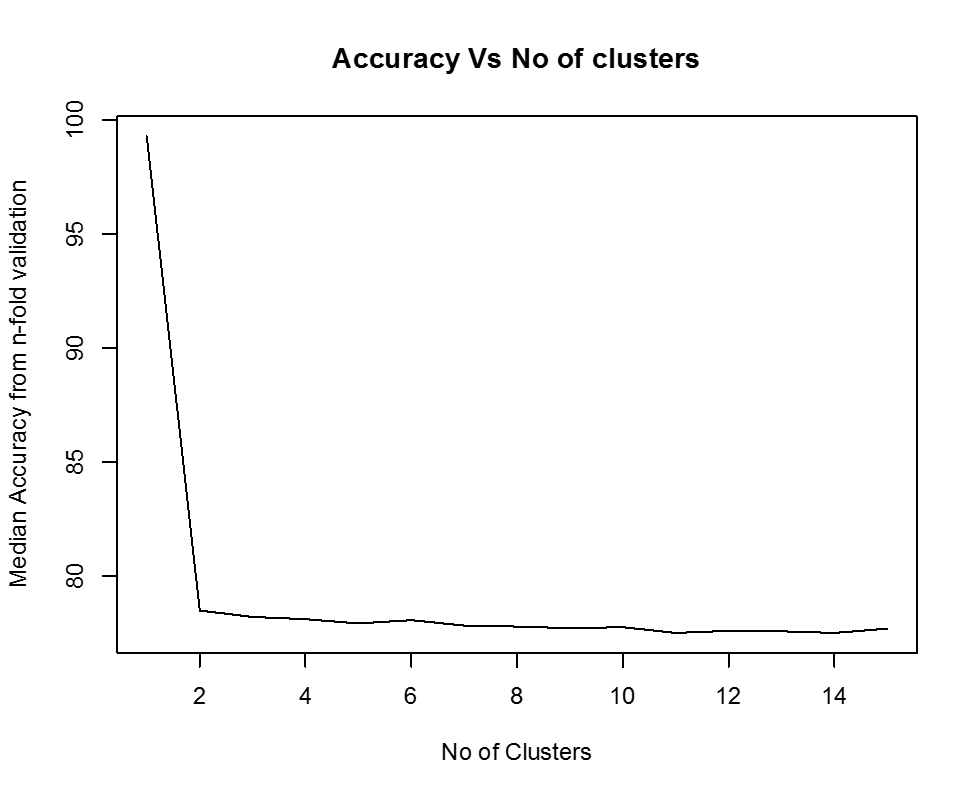


Variation of accuracy was tested by varying train dataset size in steps of ten and k-fold method where k=100. We see maximum accuracy is obtained at train set size of 260, increasing train size beyond that drops accuracy (this could be due to over fitting)

Using this train set size, we test for accuracy variation by varying clusters – We see that accuracy is maximum at 12-13 cluster. Looking at cluster means, we can make out that 12 clusters make more business sense.

**Maximum Accuracy = 83.56%**

**KNN on Human Sensing Data**

****

Inter cluster gap is not clearly distinguishable in this dataset, since accuracy of classification is more or less same for 2-14 clusters. Relatively 6 clusters differentiate the data best – Optimum accuracy is attained at training dataset 8000 but that would leave no data for test data so we would use 6000 train data points.

**Maximum Accuracy = 77.28%**

**Heart Disease dataset is better classified using KNN**