FINAL PROJECTREPORT

The main goal is to predict whether client will subscribe a term deposit or not

Data access:

Data is collected from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing  
It has a total of 41,188 records and 20 inputs from May 2008-Nov 2010

Predictor(input) and Target variables:

Predictor variables can be classified in to three types.  
1.Demographic data includes age,job,education, marital,education,default(it refers to whether he has a credit loan or not),housingloan

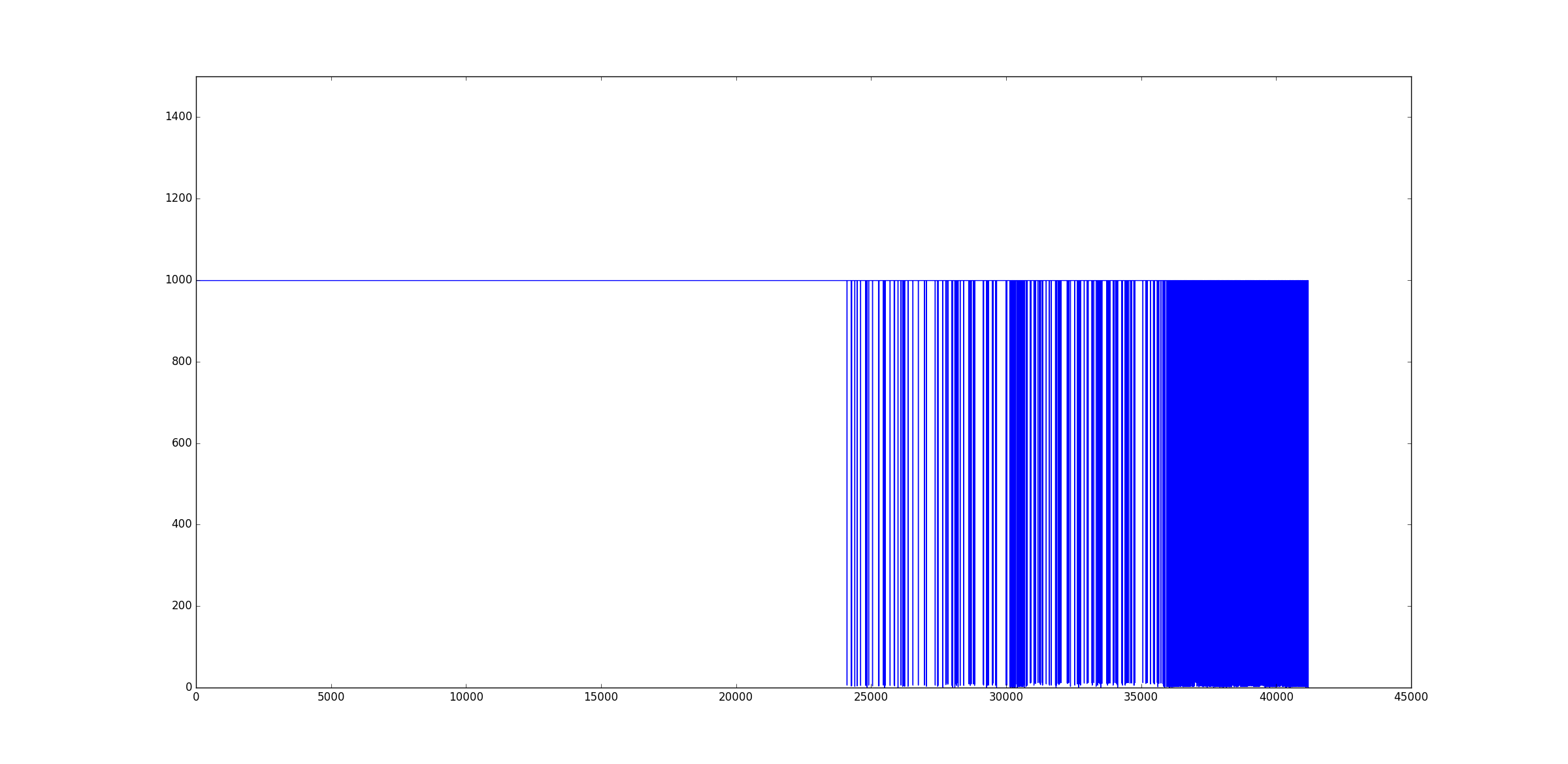
2.Data collected from previous contact which includes month,day,week,duration(cal duration during their previous contact)  
3.Social and Economic attributes includes employeevariation rate(emp.var.rate),consumerprice index(cons.price index),euribor3m indicator,nr.employed)  
Target variable is a binary variable which predicts 1 if term deposit is subscribed or zero if it is not subscribed

Data cleaning:  
 There are 20 predictor variables on which a target variable depends .There is a single target variable which corresponds to whether a term deposit is subscribed or not.

Amoong our input variables(predictor variables) there are 10 categorical and 10 numerical inputs  
 among them  
The attribute pdays refers to the total days customer last contacted

If a customer was not contacted before, they have given the value for pdata as 999  
so more than 95% of data has pvalue of 999. Only remaining data has different p value which will have negligible influence after dividing train and test data

The below plot refers to pvalue for all the records

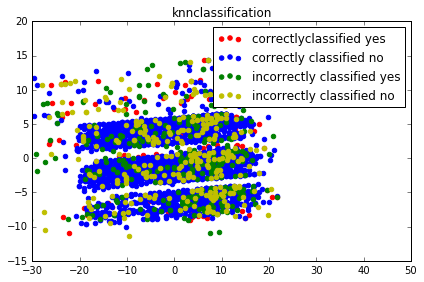


Duration also has been removed since it is mentioned in data description that it is for benchmark purpose

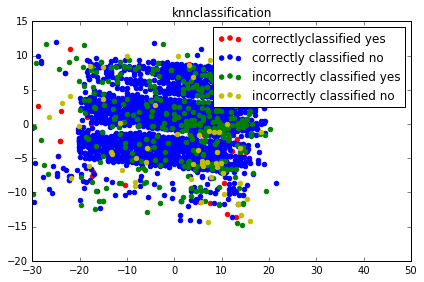
Since emp\_var\_rate has high covariance with consumer Confidence Index (cons\_conf\_index) and ConsumerPriceIndex(cons\_price\_index).so, cons\_conf\_index and cons\_price\_index has been removed

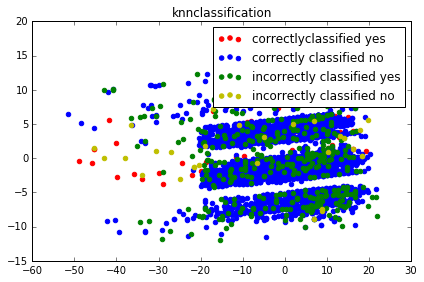
K nearest neighbor:

This plot is for N=1 nearest neighbour

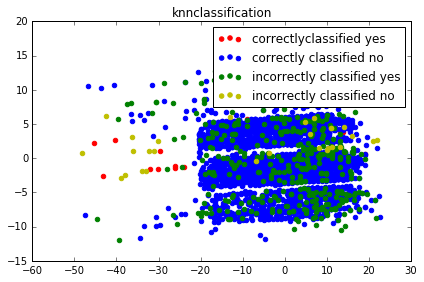


This plot is for N=10 nearest neighbour

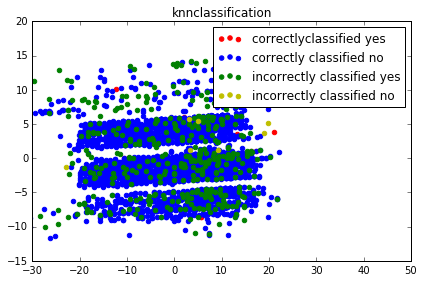




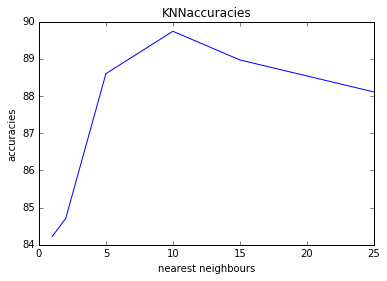
N=80



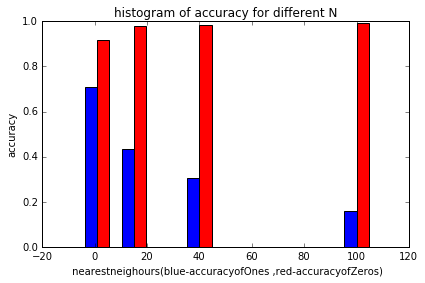
N=100



Plot for the classification accuracy for various values of N(nearest neighbours)



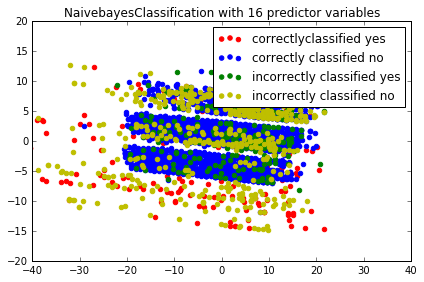
The best kvalue for this dataset is 10  
From the above graph, it can be inferred that accuracy increases with increase in nearest neighbours up to certain N,and then accuracy gradually decreases with increase in K after that.This is because the target variable has more number of zeros(no)than ones(yes). With increase in number of nearest neighbours, proability of predicting yes is gradually decreasing which is resulting in decrease in accuracy



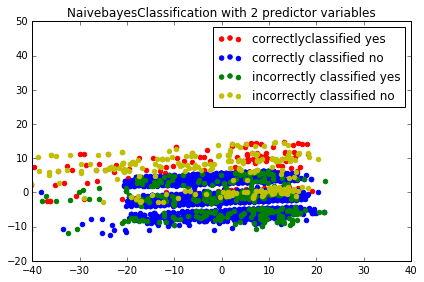
As mentioned above the accuracy of predicting ‘NO’target value decreases with increase in K value

NAIVEBAYES:

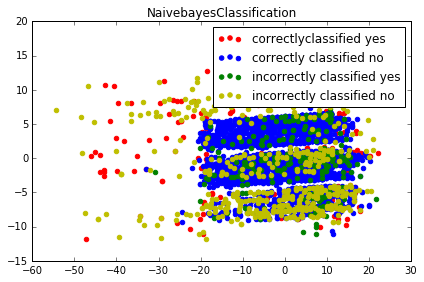
The input predictor variables is reduced to 2dimensions using PCA and then the graph is plotted for correctly and incorrectly classified zeros and ones

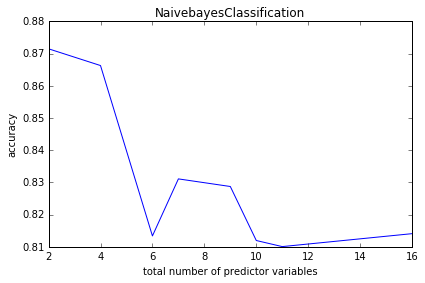


Here prediction is made by considering two highly covariant input variables with target



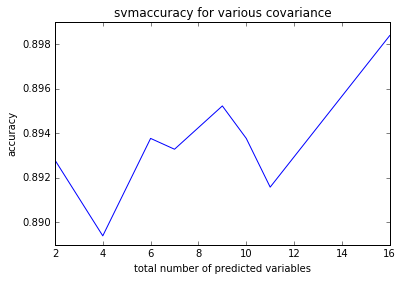
NAivebayes classification with all inout variables





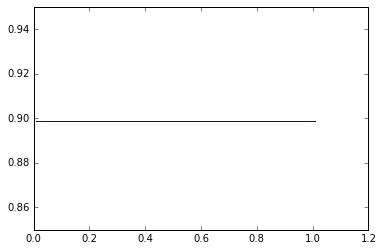
The proability of naïve bayes assumption is decreasing i.e., the naïve bayes assumption that input variables are independent is not true  
The input predictor variables are not independent

SVM:

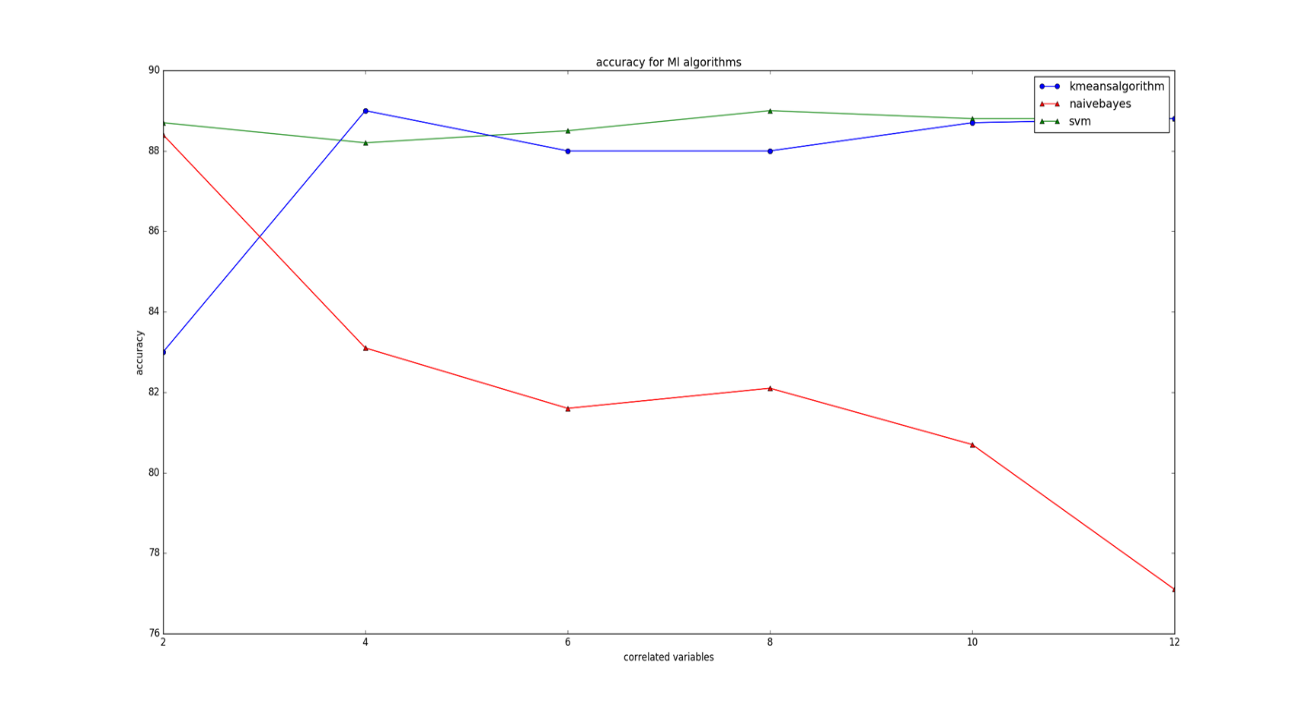


total number of predicted variables has little influence of N

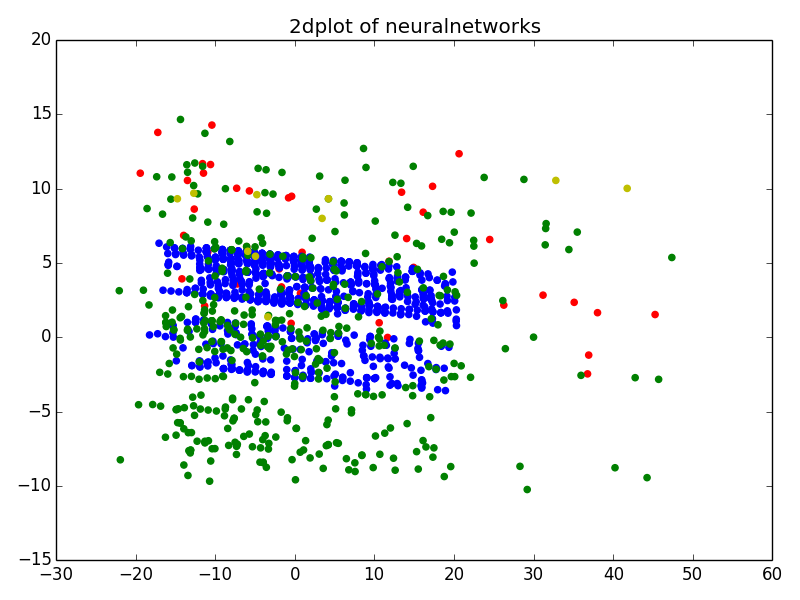
From the observation as plotted below, it is noted that the value of C has no much influence on accuracy for my data



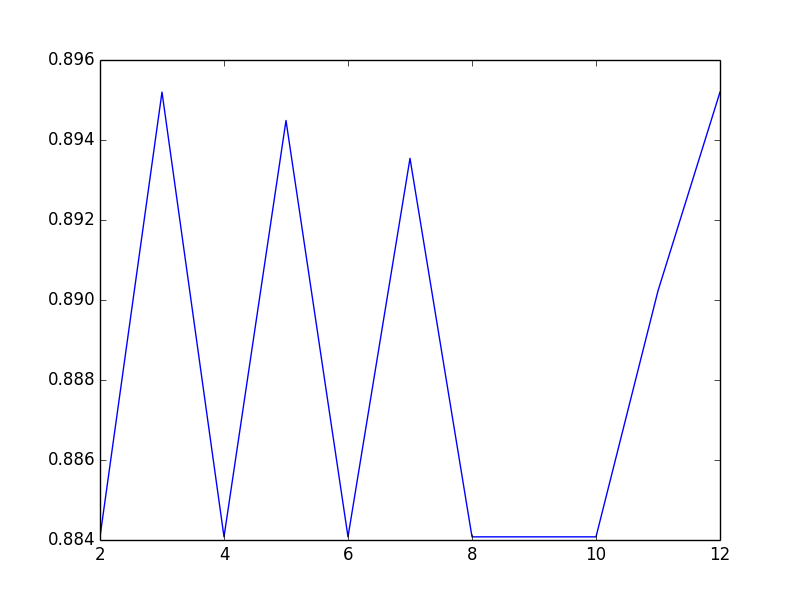
Comparing the above three algorithms for various correlated variables



From the above graph we can see that Knn and svm accuracy is increasing with increase in number of correlated variables but it is reverse in case of naïve bayes

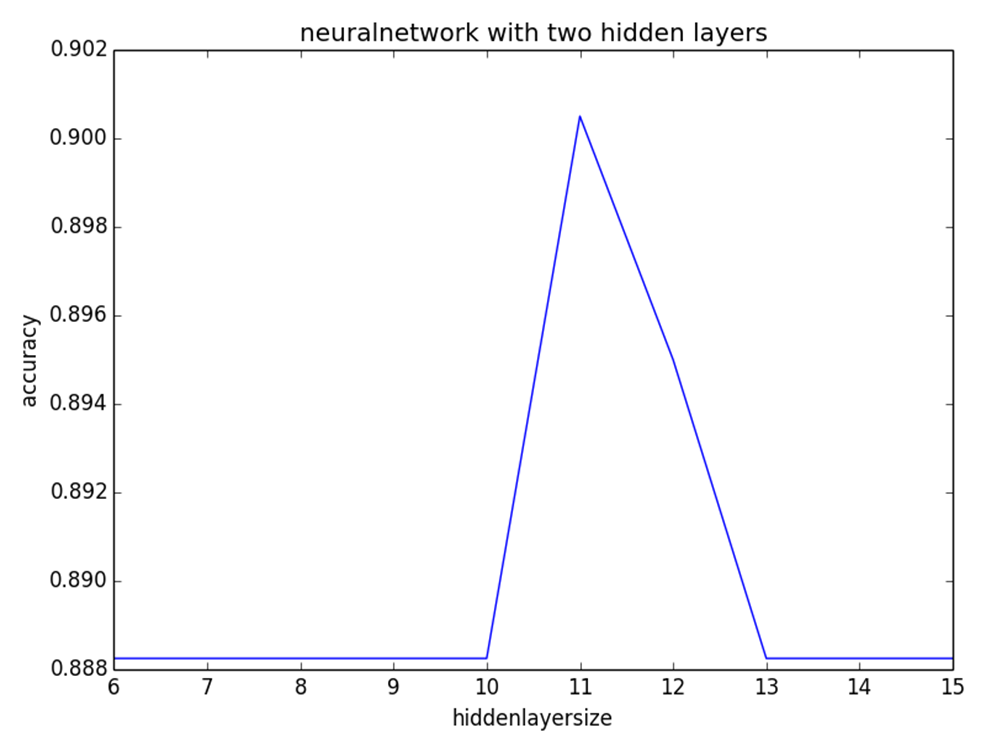
Neural networks   


with one hidden layer:

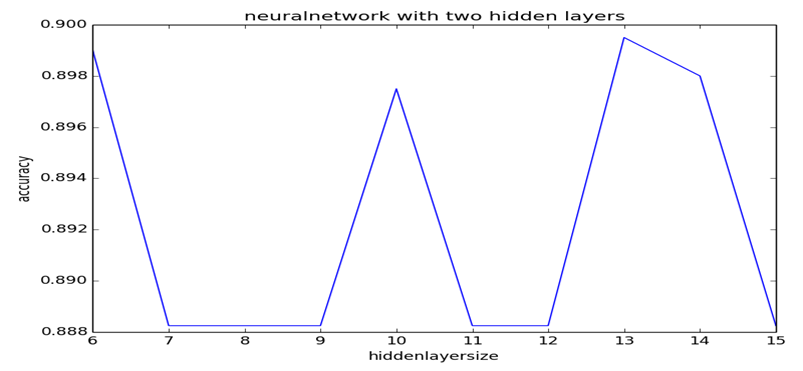


NeuralNetworks with 2hidden layers:  
The X coordinate specifies the length of first hidden layer size

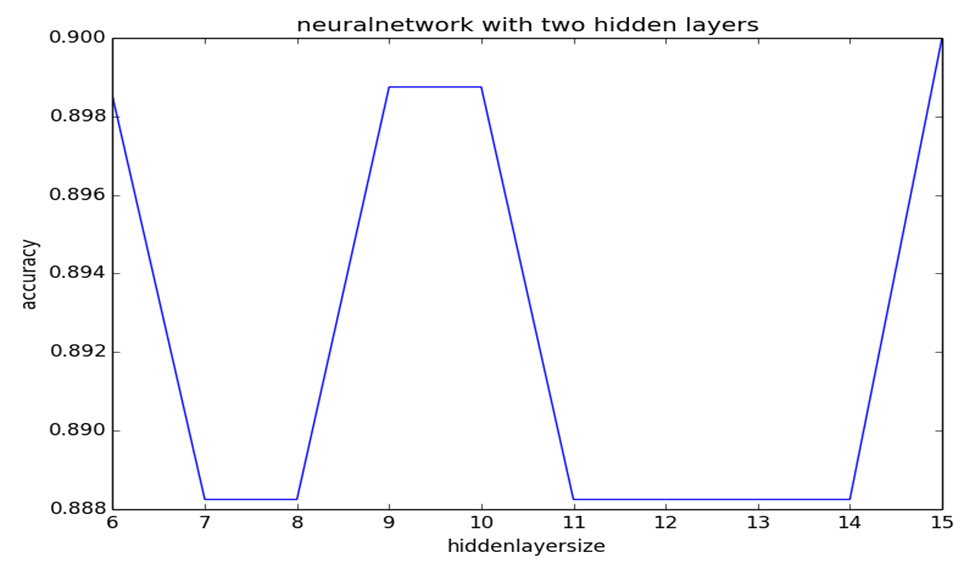
Second hiddenlayer size:2

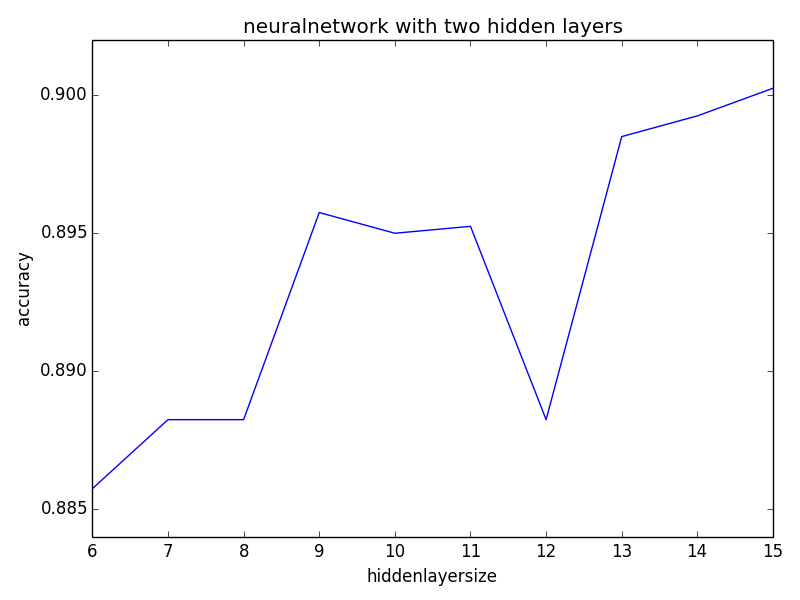


Second hiddenlayer size:3



Second hiddenlayer size:4





Second hiddenlayer size:5

Best accuracies for all Algorithms:

|  |  |
| --- | --- |
| **Algorithm** | **BestAccuracy** |
| KNN | 89.7% |
| Naïve bayes | 87.1% |
| SVM | 89.8% |
| Neural networks | 90% |