

**Breast Cancer**

**Classification**



**SUBMITTED BY-**

**SANCHITA THAKUR**

**18A3HP644**

**SUBMITTED TO-**

**PROF. SRIDHAR.V**

**September 9, 2019**

**Institute of Management Technology, Hyderabad**

**INTRODUCTION**

Breast cancer is the second most occurring cancer in the women all over the world. The number and the size of databases recording medical data are increasing rapidly. Medical data, produced from measurements, examinations, prescriptions, etc., are stored in different databases on a continuous basis. This enormous amount of data exceeds the ability of traditional methods to analyze and search for interesting patterns and information that is hidden in them. Therefore, new techniques and tools for discovering useful information in these data depositories are becoming more demanding. Here, I am going to use the logistic regression to check that how much is the probability of a particular variable for accruing the breast cancer. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

**Data Set**

The data set breast cancer which I have taken has 570 rows. The number of variables are 33 which are id, diagnosis, radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, fractal dimension mean , radius se, texture se, perimeter se, area se, smoothness se, compactness se, concave points se, symmetry se, fractal dimension se, radius worst, texture worst, perimeter worst, area worst, smoothness worst, compactness worst , concavity worst, concave points worst, symmetry worst, fractal dimension worst.

**Source of the dataset**

Kaggle is the source of the dataset.

Link- <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

**Objective**

To predict that which variable is causing the most effect on the breast cancer.

**Variable/ Data Definition**

**Independent Variable-**

Unit (mm)

1. Diagnosis- The diagnosis of breast tissues (M = malignant, B = benign)
2. Radius mean- mean of distances from center to points on the perimeter
3. Texture mean- standard deviation of gray-scale values
4. Perimeter mean- mean size of the core tumor
5. Area mean
6. Smoothness mean- mean of local variation in radius lengths
7. Compactness mean - mean of perimeter^2 / area - 1.0 (Solidity)
8. Concavity mean- mean of severity of concave portions of the contour(curved inwards)
9. concave points mean- mean for number of concave portions of the contour
10. symmetry mean – Mean of the equality
11. fractal dimension mean - mean for "coastline approximation" - 1
12. radius se - standard error for the mean of distances from center to points on the perimeter
13. texture se - standard error for standard deviation of gray-scale values
14. perimeter se -Standard error of the parameter
15. area se – Standard error of area
16. smoothness se- standard error for local variation in radius lengths
17. compactness se- standard error for perimeter^2 / area - 1.0
18. concavity se- standard error for severity of concave portions of the contour
19. concave points se- standard error for number of concave portions of the contour
20. symmetry se- Standard error of the equality
21. fractal dimension se- standard error for "coastline approximation" - 1
22. radius worst- "worst" or largest mean value for mean of distances from center to points on the perimeter
23. texture worst- "worst" or largest mean value for standard deviation of gray-scale values
24. perimeter worst
25. area worst smoothness worst- "worst" or largest mean value for local variation in radius lengths
26. compactness worst- "worst" or largest mean value for perimeter^2 / area - 1.0
27. concavity worst- "worst" or largest mean value for severity of concave portions of the contour
28. concave points worst- "worst" or largest mean value for number of concave portions of the contour
29. symmetry worst-
30. fractal dimension worst-"worst" or largest mean value for "coastline approximation" – 1
31. ID number

**Dependent Variable-** Diagnosis(Benign, Malignant)

**Machine Learning Techniques Applied**

**Logistic Regression-** The logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

**KNN model Implementation -** . KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

**Classification -** A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. Feature: A feature is an individual measurable property of a phenomenon being observed.

**Machine Learning Pipeline**

**Data Wrangling-**

Before making anything like feature selection, feature extraction and classification, firstly we start with basic data analysis. Lets look at features of data.

**There are 4 things that take my attention** 1) There is an **id** that cannot be used for classification 2) **Diagnosis** is our class label 3) **Unnamed: 32** feature includes NaN so we do not need it. 4) I do not have any idea about other feature names actually I do not need because machine learning is awesome **:)**

Therefore, drop these unnecessary features. However do not forget this is not a feature selection. This is like a browse a pub, we do not choose our drink yet !!!

Number of Benign: 357

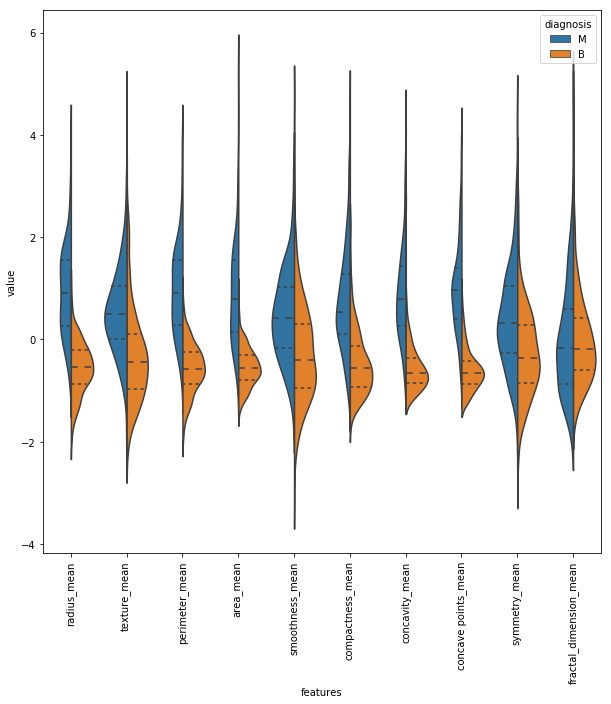
Number of Malignant : 212

****

**Data Visualization-**

1. In order to visualize data we are going to use seaborn plots that is not used in other kernels to inform you and for diversity of plots. We generally use violin plot and swarm plot. Do not forget we are not selecting feature.

Before violin and swarm plot we need to do normalization or standardization . Because differences between values of features are very high to observe on plot. I plot features in 3 group and each group includes 10 features to observe better.



**Interpretation**

Lets interpret the plot above together. For example, in **texture mean** feature, median of the Malignant and Benign looks like separated so it can be good for classification. However, in **fractal dimension mean** feature, median of the Malignant and Benign does not looks like separated so it does not gives good information for classification.

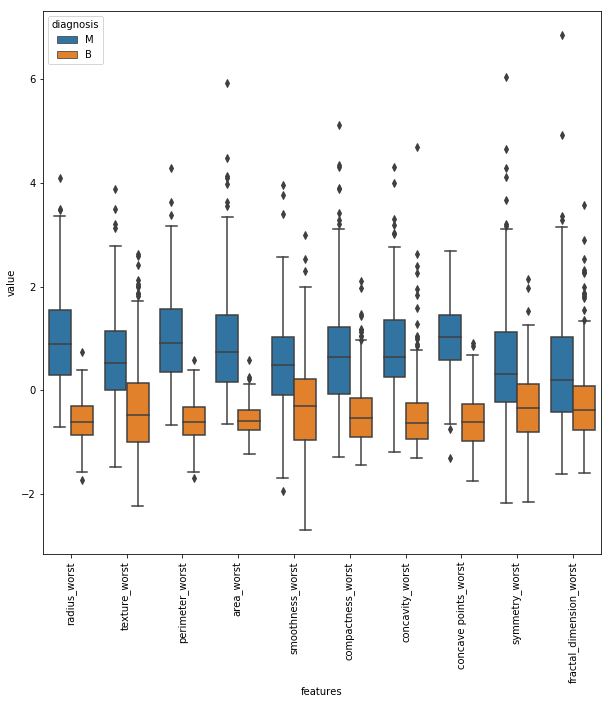
1. As an alternative of violin plot, box plot can be used

Box plots are also useful in terms of seeing outliers

I do not visualize all features with box plot

In order to show you lets have an example of box plot

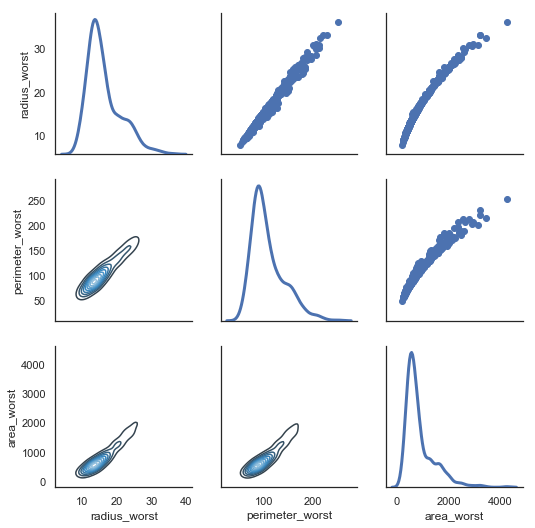
If you want, you can visualize other features as well.



**Interpretation -** Lets interpret one more thing about plot above, variable of **concavity worst** and **concave point worst** looks like similar but how can we decide whether they are correlated with each other or not. (Not always true but, basically if the features are correlated with each other we can drop one of them)

In order to compare two features deeper, lets use joint plot. Look at this in joint plot below, it is really correlated. Pearson r value is correlation value and 1 is the highest. Therefore, 0.86 is looks enough to say that they are correlated. Do not forget, we are not choosing features yet, we are just looking to have an idea about them.

1. What about three or more feature comparison ? For this purpose we can use pair grid plot. Also it seems very cool :) And we discover one more thing **radius worst**, **perimeter worst** and **area worst** are correlated as it can be seen pair grid plot. We definitely use these discoveries for feature selection.

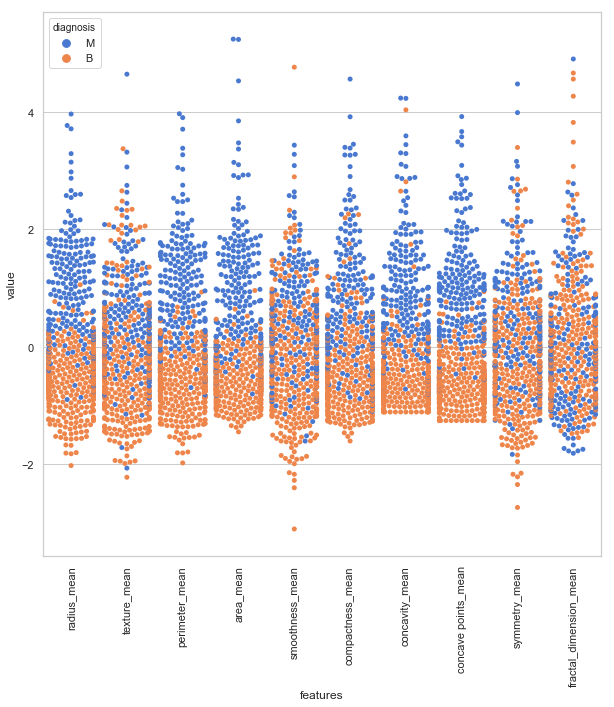


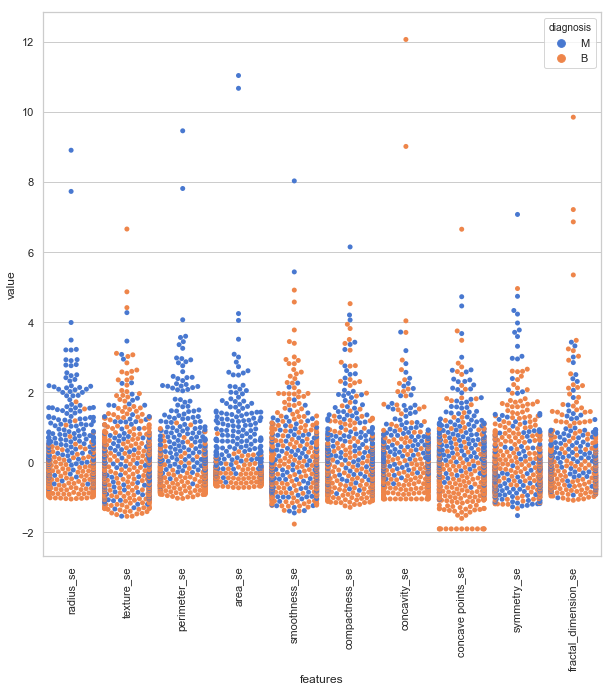
Up to this point, we make some comments and discoveries on data already. If you like what we did, I am sure swarm plot will open the pub's door :)

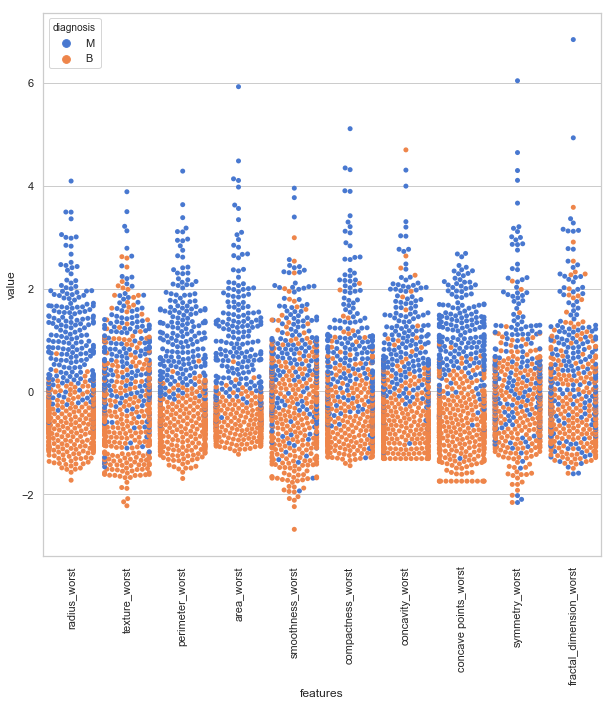
In swarm plot, I will do three part like violin plot not to make plot very complex appearance

1. They looks cool right. And you can see variance more clear. Let me ask you a question, **in these three plots which feature looks like more clear in terms of classification.** In my opinion **area worst** in last swarm plot looks like malignant and benign are separated not totally but mostly. However, **smoothness se** in swarm plot 2 looks like malignant and benign are mixed so it is hard to classify while using this feature.

**What if we want to observe all correlation between features?** Yes, you are right. The answer is heatmap that is old but powerful plot method.







They looks cool right. And you can see variance more clear. Let me ask you a question, **in these three plots which feature looks like more clear in terms of classification.** In my opinion **area worst** in last swarm plot looks like malignant and benign are separated not totally but mostly. However, **smoothness se** in swarm plot 2 looks like malignant and benign are mixed so it is hard to classify while using this feature.

**What if we want to observe all correlation between features?** Yes, you are right. The answer is heatmap that is old but powerful plot method.

**Model Building**

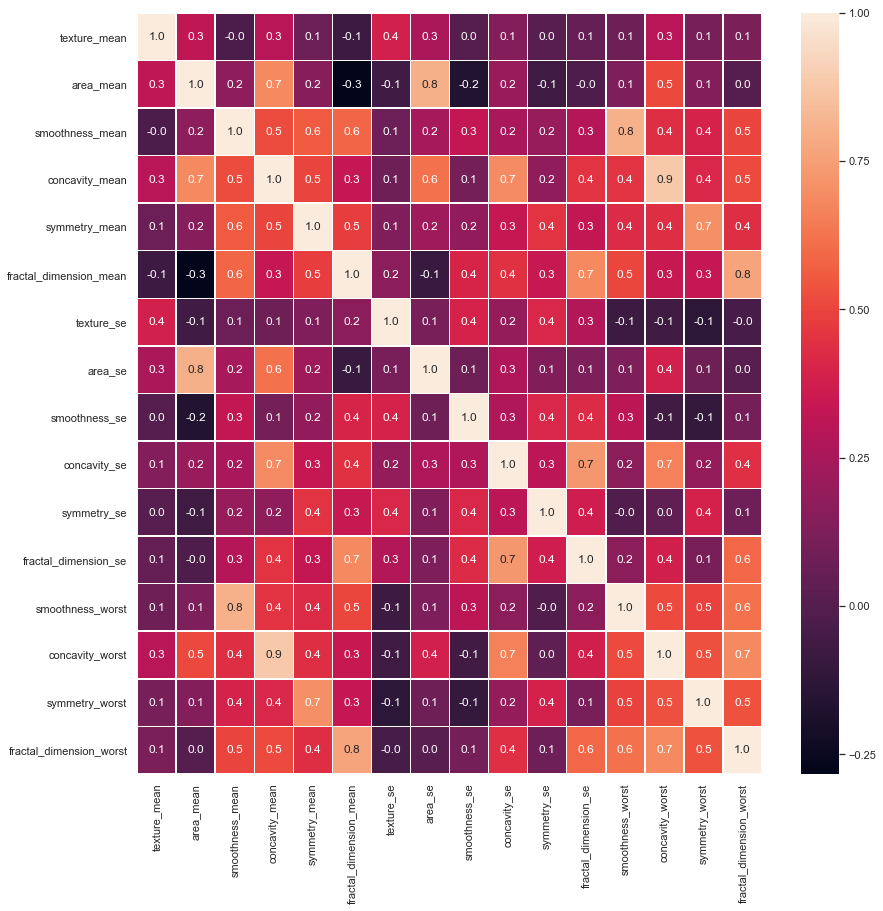
**Feature Selection and Random Forest Classification**

In this part we will select feature with different methods that are feature selection with correlation, univariate feature selection, recursive feature elimination (RFE), recursive feature elimination with cross validation (RFECV) and tree based feature selection. We will use random forest classification in order to train our model and predict.

**1) Feature selection with correlation and random forest classification**

As it can be seen in map heat figure **radius mean, perimeter mean and area mean** are correlated with each other so we will use only **area mean**. If you ask how i choose **area mean** as a feature to use, well actually there is no correct answer, I just look at swarm plots and **area mean** looks like clear for me but we cannot make exact separation among other correlated features without trying. So lets find other correlated features and look accuracy with random forest classifier.

**Compactness mean, concavity mean and concave points mean** are correlated with each other. Therefore I only choose **concavity mean**. Apart from these, **radius se, perimeter se and area se** are correlated and I only use **area se**. **radius worst, perimeter worst and area worst** are correlated so I use **area worst**. **Compactness worst, concavity worst and concave points worst** so I use **concavity worst**. **Compactness se, concavity se and concave points se** so I use **concavity se**. **texture mean and texture worst are correlated** and I use **texture mean**. **Area worst and area mean** are correlated, I use **area mean**.



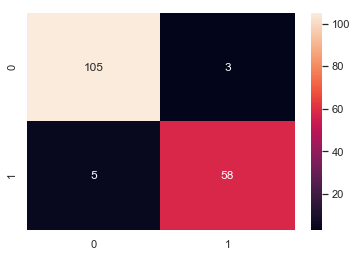
After drop correlated features, as it can be seen in below correlation matrix, there are no more correlated features. Actually, I know and you see there is correlation value 0.9 but lets see together what happen if we do not drop it.

Well, we choose our features but **did we choose correctly ?** Lets use random forest and find accuracy according to chosen features.

**Accuracy- TP+TN/P+N**

**= 105+58/105+58+5+3 = .95**

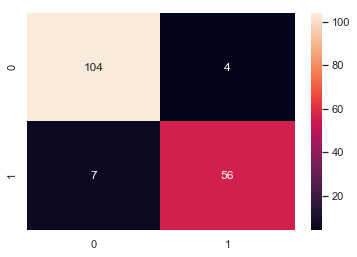
**= 95%**



Accuracy is almost 95% and as it can be seen in confusion matrix, we make few wrong prediction. Now lets see other feature selection methods to find better results.

## 2) Univariate feature selection and random forest classification[¶](http://localhost:8888/notebooks/Downloads/Feature%20Selection%20and%20Data%20Visualization.ipynb#2)-Univariate-feature-selection-and-random-forest-classification)

In this method we need to choose how many features we will use. For example, will k (number of features) be 5 or 10 or 15? The answer is only trying or intuitively. I do not try all combinations but I only choose k = 5 and find best 5 features.



Accuracy is almost 96% and as it can be seen in confusion matrix, we make few wrong prediction. What we did up to now is that we choose features according to correlation matrix and according to select k Best method. Although we use 5 features in select k Best method accuracies look similar. Now lets see other feature selection methods to find better results.

## 3) Recursive feature elimination (RFE) with random forest

Basically, it uses one of the classification methods (random forest in our example), assign weights to each of features. Whose absolute weights are the smallest are pruned from the current set features. That procedure is recursively repeated on the pruned set until the desired number of features

Like previous method, we will use 5 features. However, which 5 features will we use ? We will choose them with RFE method.

Chosen 5 best features by RFE is **texture mean, area mean, concavity mean, areas, concavity worst**. They are exactly similar with previous (select k Best) method. Therefore we do not need to calculate accuracy again. Shortly, we can say that we make good feature selection with RFE and select k Best methods. However as you can see there is a problem, okay I except we find best 5 feature with two different method and these features are same but why it is **5**. Maybe if we use best 2 or best 15 feature we will have better accuracy. Therefore lets see how many feature we need to use with RFE cv method.

**4) Recursive feature elimination with cross validation and random forest classification**

Now we will not only **find best features** but we also find **how many features do we need** for best accuracy.

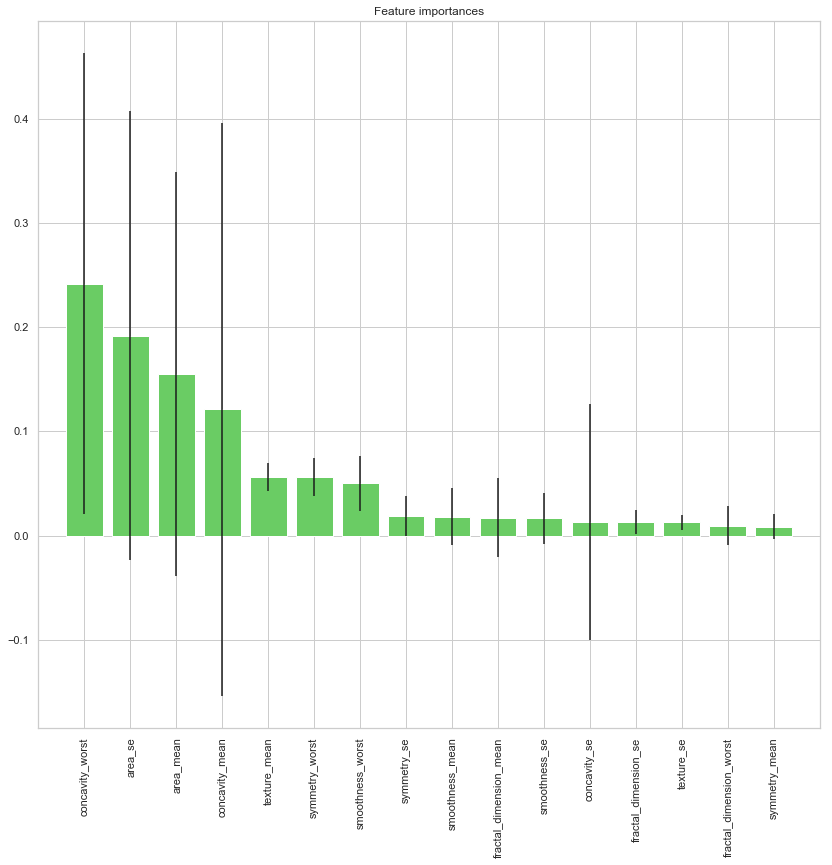
Finally, we find best 11 features that are **texture mean, area mean, concavity mean, texture se, area se, concavity se, symmetry se, smoothness worst, concavity worst, symmetry worst and fractal dimension worst** for best classification. Lets look at best accuracy with plot.

## ****C:\Users\sanchita\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\4E71E401.tmp****

Lets look at what we did up to this point. Lets accept that guys this data is very easy to classification. However, our first purpose is actually not finding good accuracy. Our purpose is learning how to make **feature selection and understanding data.** Then last make our last feature selection method.

## 5) Tree based feature selection and random forest classification

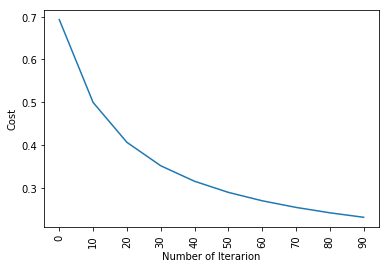
In random forest classification method there is a **feature importance’s** attributes that is the feature importance’s (the higher, the more important the feature). **!!! To use feature importance method, in training data there should not be correlated features. Random forest choose randomly at each iteration, therefore sequence of feature importance list can change.**

****

As you can seen in plot above, after 5 best features importance of features decrease. Therefore we can focus these 5 features. As I sad before, I give importance to understand features and find best of them.

**Logistic Regression**

The logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

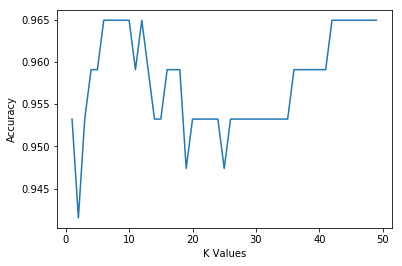
****

train accuracy: 94.28571428571429 %

test accuracy: 95.6140350877193 %

**KNN Implementation**

**KNN** is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. Just for reference, this is “where” **KNN** is positioned in the algorithm list of sci kit learn.

****

95% accuracy

**Model Evaluation**

## Univariate feature selection and random forest classification technique is the best technique as we can see in the model building. It is giving the 96% accuracy which is the highest as compared to others. The logistic regression is also the best it has 96 % accuracy as well.

# Conclusion

Shortly, I tried to show importance of feature selection and data visualization. Default data includes 33 feature but after feature selection we drop this number from 33 to 5 with accuracy 95%. In this kernel we just tried basic things, I am sure with these data visualization and feature selection methods, you can easily exceed the % 95 accuracy. Maybe you can use other classification methods.