**CLASSIFICATION OF DIFFERENT STAGES OF RENAL CLEAR CELL CARCINOMA USING CT SCAN IMAGES**

**POST GRADUATE DIPLOMA IN**

**STATISTICAL METHODS AND ANALYTICS**

****

**Submitted**

**by**

**Nabajyoti Pathak**

**DST-16/17-08**

**INDIAN STATISTICAL INSTITUTE**

Date: .............

**CERTIFICATE**

This is to certify Mr./Ms. ......................................... has done the project under my supervision and guidance (from ...................... to ..................). This is an original project report based on work carried out by him / her in partial fulfilment and requirement of the Post Graduate Diploma in Statistical Methods and Analytics programme of the Indian Statistical Institute, North-East Centre, Tezpur, Assam.

**(Name of the faculty)**

**ACKNOWLEDGEMENT**

It gives me immense pleasure and joy to express my heart full gratitude to my project supervisor Dr.Sanjit Maitra, Indian Statistical Institute, North-East Centre, Tezpur. Without his constant inspiration and encouragement, the present work would not have been possible. I would like to express my heartiest thanks to faculty members of Indian Statistical Institute, North-East centre, Tezpur, for their encouragement. Finally I thank my parents and friends, for their support and for being my constant source of inspiration.

**(Name of the student)**

**CONTENTS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **i) Certificate** | 2 | | **ii) Acknowledgements** | 3 | | **INTRODUCTION** |  | | Method Of Study | 7 | | Research Problem | 7 | | Data Source | 9 | | Description of data | 9 | | Statistical methods to be used | 9 | | **METHODOLOGY** |  | | Cleaning the Data | 16 | | Data Transformation | 18 | | Feature extraction | 19 | | Exploratory data analysis | 23 | | **RESULTS** |  | | Modelling using Decison Tree | 27 | | Modelling using Naive Bayes | 30 | | Modelling using KNN | 31 | | Comparing Naive Bayes and KNN classifier | 32 | | Modelling using ANN | 33 | | Summary of the model | 37 | | **CONCLUSION** | 38 | | Implications | 38 | | **REFERENCES** | 39 | |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Staging of renal cell carcinoma | 8 |
| CT Scan images of Renal cell carcinoma | 8 |
| Sample CT scan images of four different patients downloaded from TCIA | 16 |
| Sample tumour images | 17 |
| Histogram containing intensity values of 12 stage 1 patients before slicing | 18 |
| Histogram containing intensity values of 12 stage 1 patients after slicing | 19 |
| Box plot of various features(Size and Statistical measures of intensity) | 23 |
| Decision tree plot | 29 |
| Naive Bayes plot | 30 |
| plot for PCA | 33 |
| Final architecture of ANN model | 34 |
| Summary of the model | 37 |

**LIST OF TABLES**

|  |  |
| --- | --- |
| TNM staging system | 7 |
| Image statistics | 9 |
| Size related Features | 20 |
| Statistical measures of intensity values | 22 |
| Table for Decision tree | 28 |
| Confusion matrix for decision tree | 29 |
| Confusion matrix for Bayes classifier | 31 |
| Confusion matrix for KNN classifier | 32 |
| Comparing Naive Bayes and KNN classifier | 32 |
| Table for ANN classifier | 35 |

**METHOD OF STUDY**

Classification of Stages of renal clear cell carcinoma using CT Scan Images by applying various Machine Learning techniques.

**RESEARCH PROBLEM**

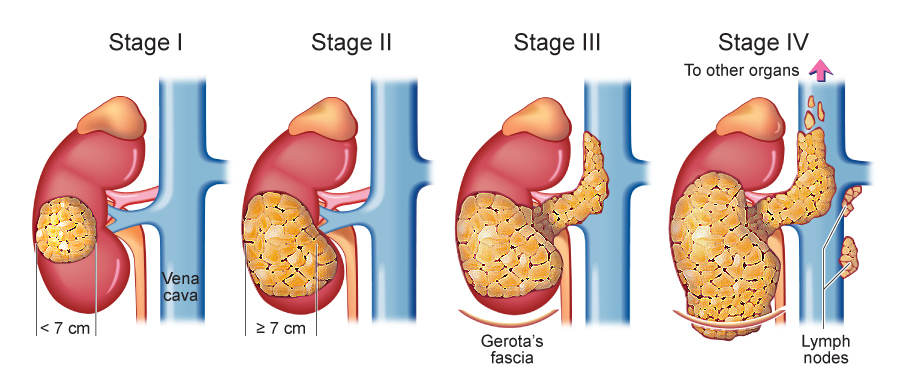
**Renal cell carcinoma** (**RCC**) is a [kidney cancer](https://en.wikipedia.org/wiki/Kidney_cancer) that originates in the lining of the [proximalconvolutedtubule](https://en.wikipedia.org/wiki/Proximal_tubule), a part of the very small tubes in the kidney that transport waste molecules from the blood to the urine. RCC is the most common type of kidney cancer in adults, responsible for approximately 90–95% of cases.

Initial treatment is most commonly either partial or complete removal of the affected kidney(s).In case the cancer has not metastasised or spread deeper into the tissues of the kidney, the 5-year survival rate is 65–90%, but this is significantly reduced when the cancer has spread.

The body is remarkably good at hiding the symptoms and as a result people with RCC often have advanced disease by the time it is discovered.  When RCC metastasises, it most commonly spreads to the [lymph nodes](https://en.wikipedia.org/wiki/Lymph_nodes), [lungs](https://en.wikipedia.org/wiki/Lungs), [liver](https://en.wikipedia.org/wiki/Liver), [adrenal glands](https://en.wikipedia.org/wiki/Adrenal_glands), [brain](https://en.wikipedia.org/wiki/Brain) or bones. [Immunotherapy](https://en.wikipedia.org/wiki/Immunomodulatory) and [targeted therapy](https://en.wikipedia.org/wiki/Targeted_therapy) have improved the outlook for metastatic RCC.

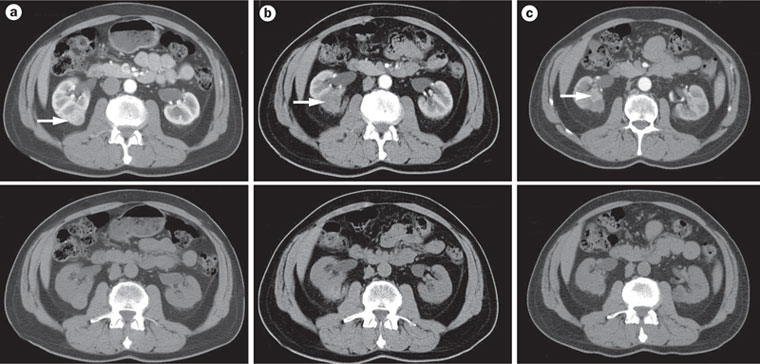
The [**staging**](https://en.wikipedia.org/wiki/Cancer_staging) of renal cell carcinoma is the most important factor in predicting its prognosis. Staging follows the [**TNM staging system**](https://en.wikipedia.org/wiki/TNM_staging_system), where the size and extent of the tumour (T), involvement of lymph nodes (N) and metastases (M) are classified separately. It is grouped into four stages (I-IV) by 1997 revision of American Joint Committee of Cancer as described below:

|  |  |
| --- | --- |
| **Stage I** | Tumour of a diameter of 7 cm (approx. 23⁄4 inches) or smaller, and limited to the kidney. No lymph node involvement or metastases to distant organs. |
| **Stage II** | Tumour larger than 7.0 cm but still limited to the kidney. No lymph node involvement or metastases to distant organs. |
| **Stage III** any of the following | Tumour of any size with involvement of a nearby lymph node but no metastases to distant organs. Tumour of this stage may be with or without spread to fatty tissue around the kidney, with or without spread into the large veins leading from the kidney to the heart. |
| Tumour with spread to fatty tissue around the kidney and/or spread into the large veins leading from the kidney to the heart, but without spread to any lymph nodes or other organs. |
| **Stage IV** any of the following | Tumour that has spread directly through the fatty tissue and the fascia ligament-like tissue that surrounds the kidney. |
| Involvement of more than one lymph node near the kidney |
| Involvement of any lymph node not near the kidney |
| Distant metastases, such as in the lungs, bone, or brain. |



Staging of renal cell carcinoma

**Contrast-enhanced**[**computed tomography**](https://en.wikipedia.org/wiki/Computed_tomography)**(CT)** scanning is routinely used to determine the stage of the renal cell carcinoma in the [abdominal](https://en.wikipedia.org/wiki/Abdominal) and [pelvic](https://en.wikipedia.org/wiki/Pelvic) regions. CT scans have the potential to distinguish solid masses from cystic masses and may provide information on the localization, stage or spread of the cancer to other organs of the patient. Key parts of the human body which are examined for [metastatic](https://en.wikipedia.org/wiki/Metastatic) involvement of renal cell carcinoma may include the [renal vein](https://en.wikipedia.org/wiki/Renal_vein), [lymph node](https://en.wikipedia.org/wiki/Lymph_node) and the involvement of the [inferior vena cava](https://en.wikipedia.org/wiki/Inferior_vena_cava). According to a study conducted by Sauk et al., multidetector CT imaging characteristics have applications in diagnosing patients with clear renal cell carcinoma by depicting the differences of these cells at the cytogenic level.



**CT SCAN IMAGES OF CLEAR RENAL CELL CARINOMA**

Our goal is to classify the various stages of renal cell carcinoma based on CT scan images and various machine learning techniques.

**DATA SOURCE**-

Data is collected from the Cancer Image Archive (TCIA).TCIA is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. The data are organized as “Collections”, typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc.) or research focus. Digital Imaging and Communications in Medicine (**DICOM**)is the primary file format used by TCIA for image storage. Supporting data related to the images such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available.

**DESCRIPTION OF DATA**-

The data consists of various CT scan images of patients diagnosed with clear renal cell carcinoma. The images are further grouped into four stages i.e. Stage1 to Stage4.

| **Image Statistics** |  |
| --- | --- |
| Modalities | CT |
| Number of Patients | 267 |
| Number of Studies | 439 |
| Number of Series | 2,654 |
| Number of Images | 192,581 |
| Images Size (GB) | 91.6 |

**STATISTICAL METHODS TO BE USED**-

Features need to be extracted from the images and based on the features; various supervised learning techniques will be applied to classify the images into four stages. After classification comparison of various techniques will be done to see which technique provides the better result in terms of predicting the stages from CT scan images.

The following list of M/L techniques have been applied in our study-

DECISION TREE-

A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision.

Decision trees can be drawn by hand or created with a graphics program or specialized software. Informally, decision trees are useful for focusing discussion when a group must make a decision. Programmatically, they can be used to assign monetary/time or other values to possible outcomes so that decisions can be automated. Decision tree software is used in data mining to simplify complex strategic challenges and evaluate the cost-effectiveness of research and business decisions. Variables in a decision tree are usually represented by circles.

Example->





Play golf = YES

NAÏVE BAYES CLASSIFIER

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

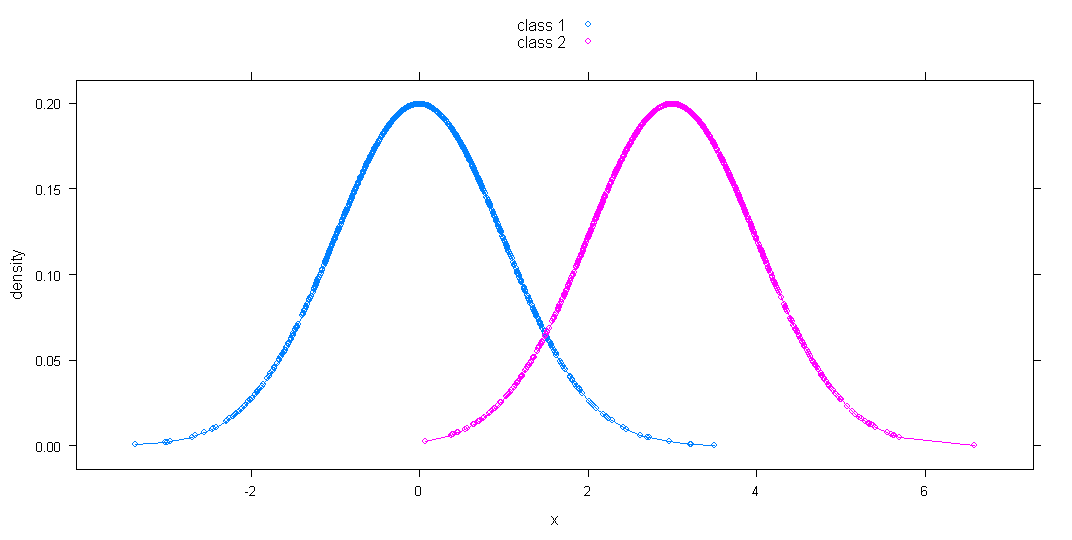
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



Above,

* P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
* P(c) is the prior probability of class.
* P(x|c) is the likelihood which is the probability of predictor given class.
* P(x) is the prior probability of predictor.

EXAMPLE->



Plot representing the Gaussian distribution of two different classes.

From the above plot we can easily see that the likelihood of “x” lying in the range of (-∞,1.5)

is more in class “C1” than “C2”.

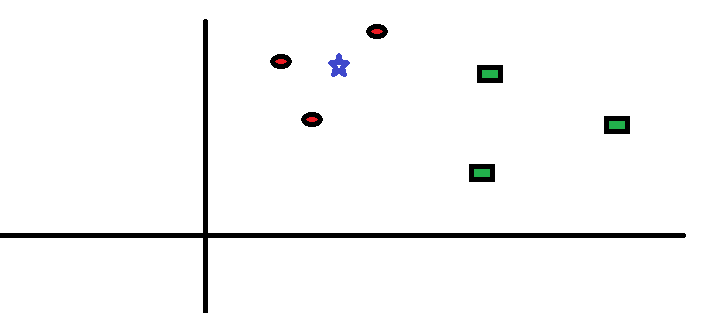
K-NEAREST NEIGHBOUR

In pattern recognition, the k-nearest neighbor’s algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

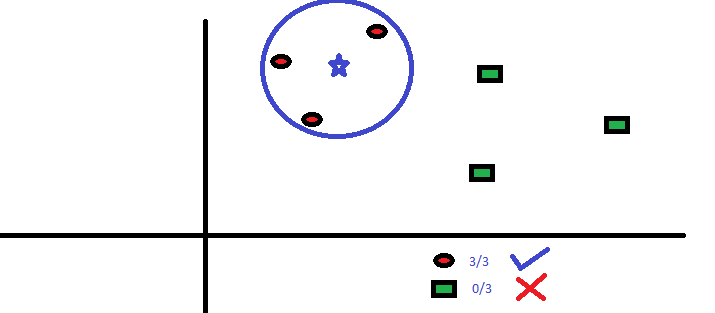
In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

EXAMPLE->

Let’s take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS):



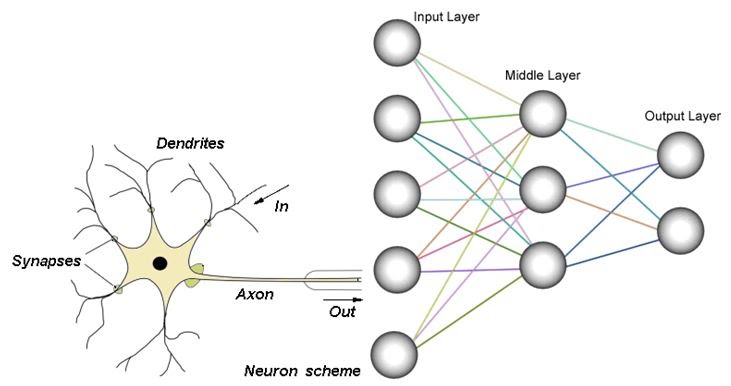
Scenario1: You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The “K” is KNN algorithm is the nearest neighbors we wish to take vote from. Let’s say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane. Refer to following diagram for more details:



scenario2: The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.



Advantages-

* A neural network can perform tasks that a linear program cannot.
* When an element of the neural network fails, it can continue without any problem by their parallel nature.
* A neural network learns and does not need to be reprogrammed.
* It can be implemented in any application.
* It can be implemented without any problem.

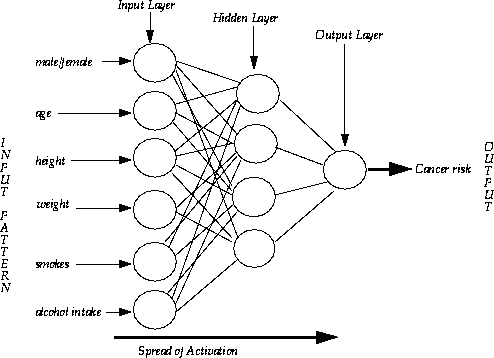
Disadvantages-

* The neural network needs training to operate.
* The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated.
* Requires high processing time for large neural networks.

An ANN is typically defined by three types of parameters-

* The interconnection pattern between different layers of neurons
* The learning process for updating the weights of the interconnections
* The activation function that converts a neuron's weighted input to its output activation.

EXAMPLE->Determine the risk of cancer based on certain inputs.

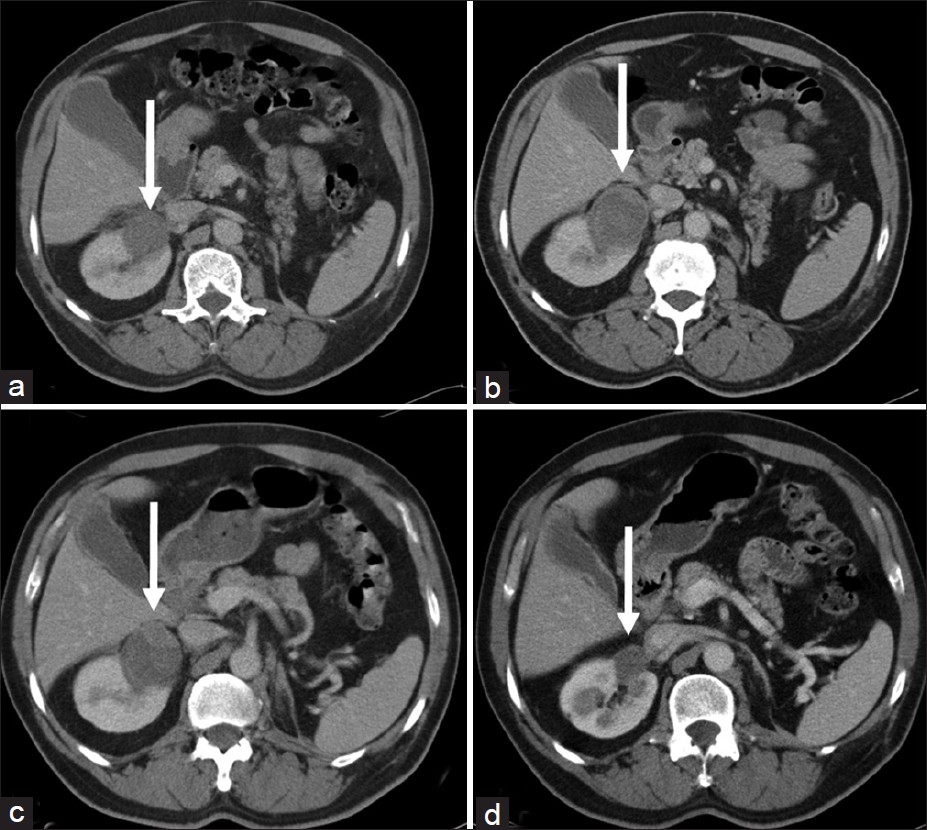


**CLEANING THE DATA**

The data contained CT scan images of 210 patients grouped into various stages. After thorough inspection it was found that only 100 of them showed clear images of the tumor.

The no of images for the different stages are given below-

|  |  |
| --- | --- |
| **STAGES** | **NO OF IMAGES** |
| **Stage1** | **32** |
| **Stage2** | **21** |
| **Stage3** | **24** |
| **Stage4** | **23** |



**Sample CT scan images of four different patients downloaded from TCIA**

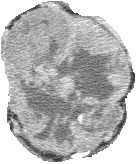
In the above images the tumor can be easily visualized as indicated by the arrow. The tumor part was cropped for feature extraction. Some of the sample images are given below.

C:\Users\lab\Desktop\project\finalimages\stage1\1.jpgC:\Users\lab\Desktop\project\finalimages\stage1\32.jpg

**Stage1**



**Stage2**



**Stage3**

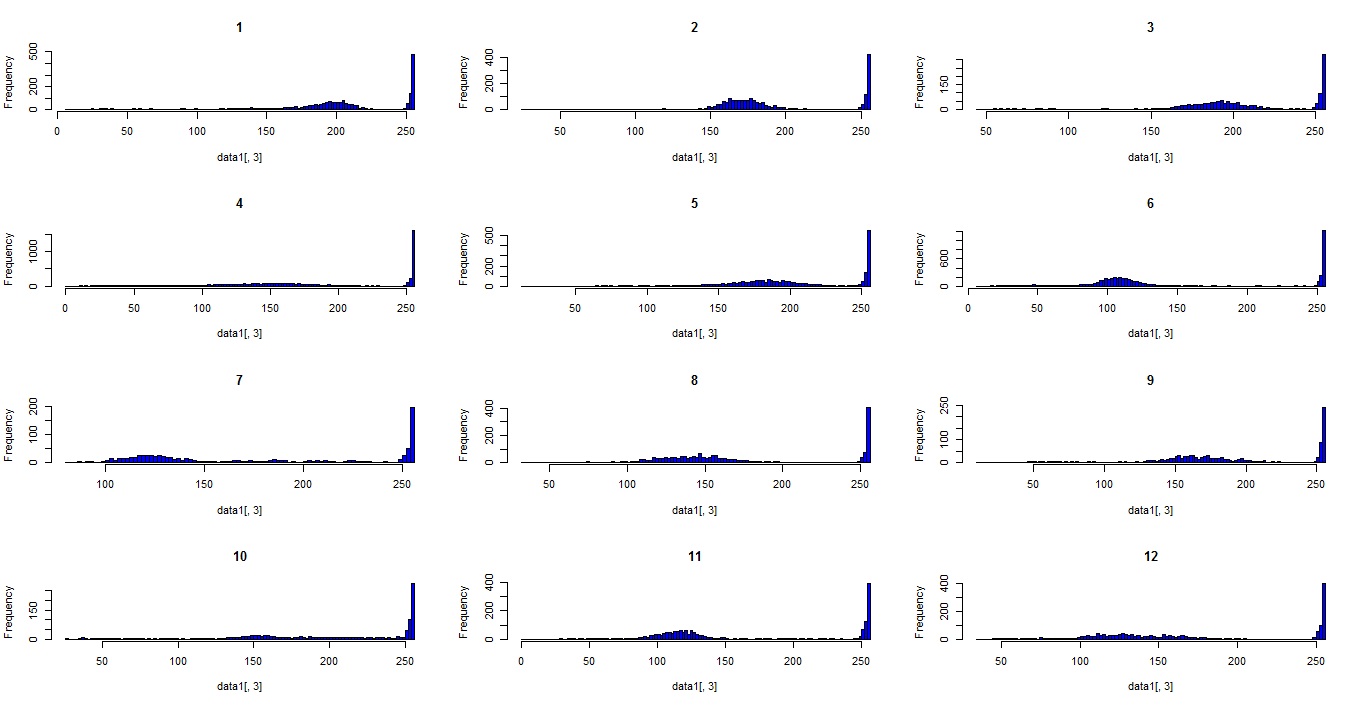


**Stage4**

#Remark- In the above images we can see that “Size” is one of the important features for classifying the images.

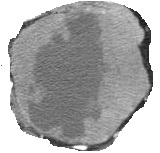
**DATA TRANSFORMATION**

The pixel values of each of these cropped images were extracted using the “OPEN CV2” module in python and histogram analysis was done on the new data to threshold the pixel values having white intensity. The intensity values range from 0 to 255(black to white)



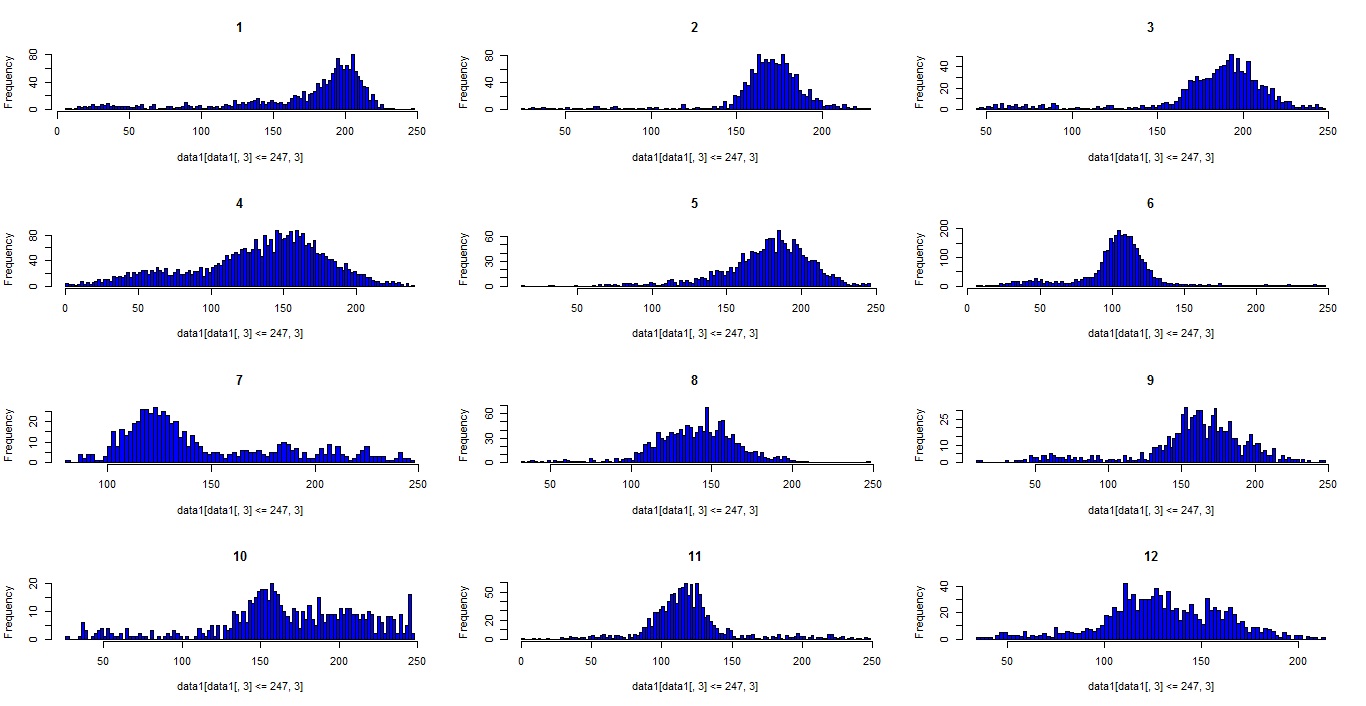
**HISTOGRAM CONTAINING INTENSITY VALUES OF 12 STAGE1 PATIENTS BEFORE SLICING**

The histogram shows a peak in the right tail indicating the white intensity values. This pixels are the white background covering the tumor which need to be removed from the data.



After studying the histogram, thresholding of pixels having intensity values “247”

and above was done to remove the white boundary surrounding the tumor.



**HISTOGRAM CONTAINING INTENSITY VALUES OF 12 STAGE1 PATIENTS AFTER SLICING**

#REMARKS- All of the above histogram indicates a certain distribution of pixel intensity values.

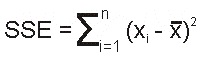
**FEATURE EXTRACTION**-

As stated above “Size” is one of the important features in classifying the images. So features pertaining to size were extracted as given below-

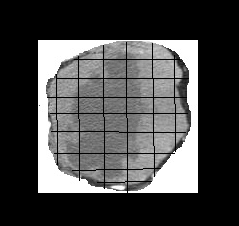
1. No of pixels in the image.

2. Sum of Squared error using the pixel locations(x,y)

3. Radius of the Tumor.

 Xi=i’th pixel location value

=mean of all the pixel location values



Radius

Pixels

The table below indicates the features as stated above –





Only size related feature is not enough to classify the stages as already mentioned above for stages 2, 3 and 4 size is not helpful to distinguish them. So various statistical properties related to pixel intensity values were extracted to see if it helps in classifying the stages.

The following statistical measures were taken to extract features related to pixel intensity values-

1. MEAN 2.MEDIAN 3.MODE 4.STANDARD DEVIATION 5.SKEWNESS

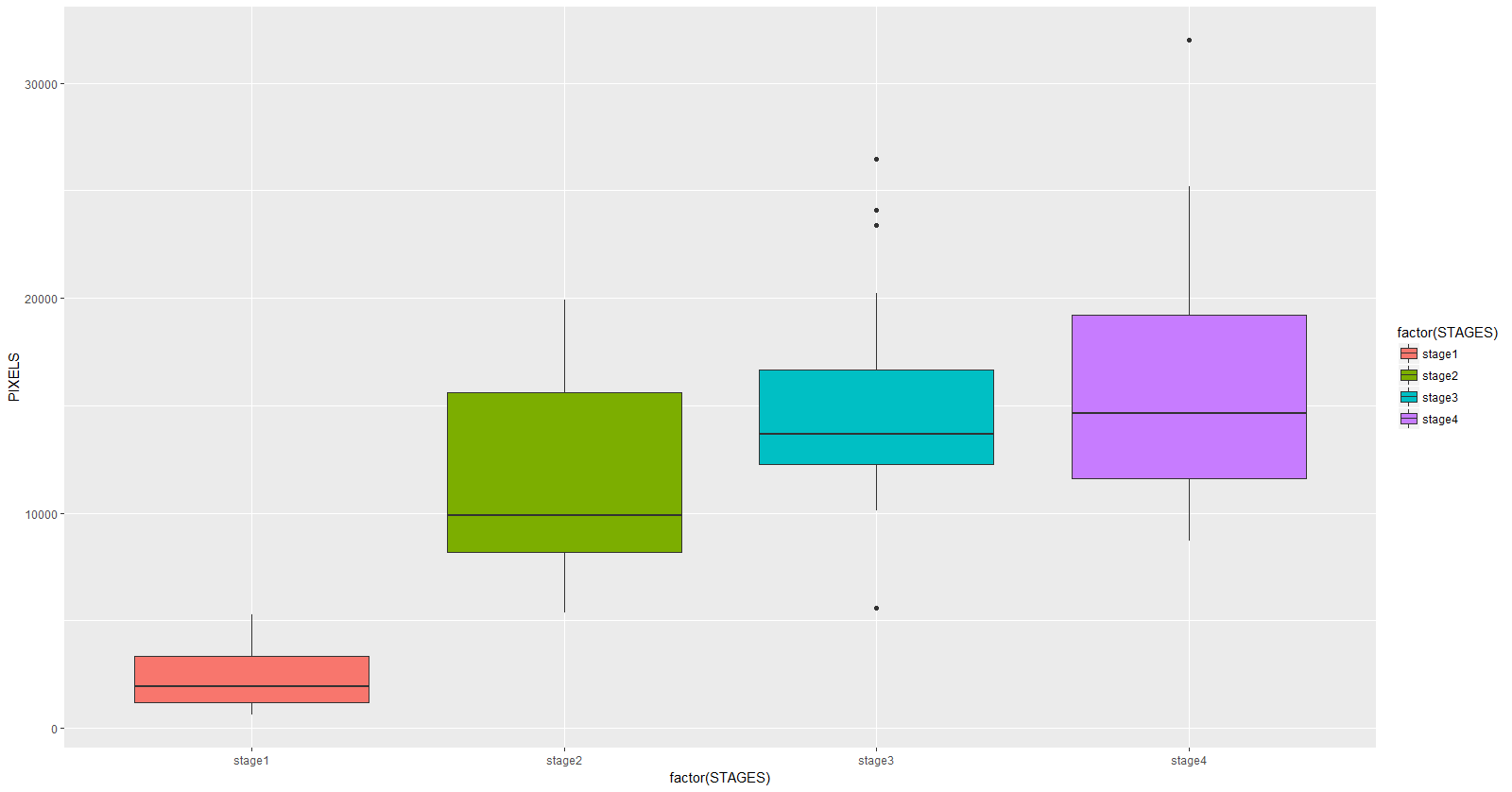
6. KURTOSIS



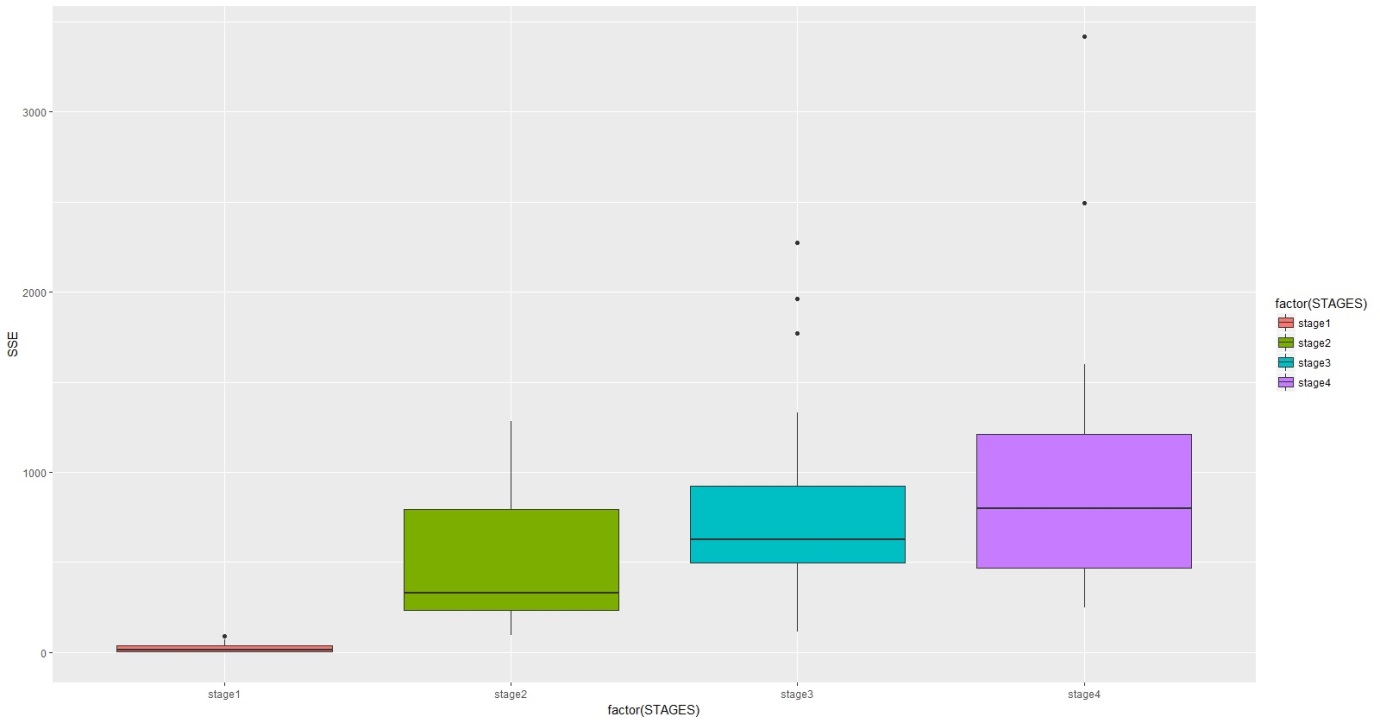
#Remark-Some of the features was standardized using Range(x) function.

**EXPLORATORY DATA ANALYSIS**

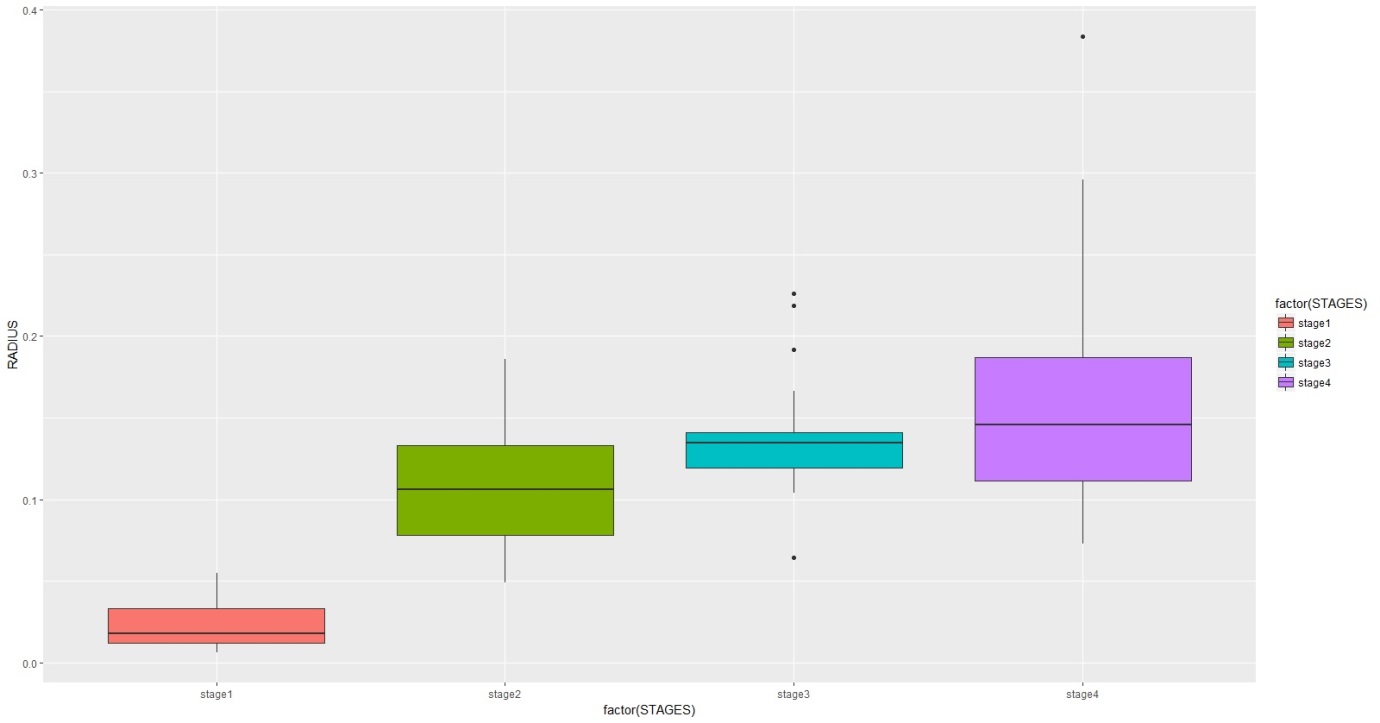
Box plots of various stages using the above features were done using R for exploratory data analysis.



**BOX PLOT CONTAINING NO OF PIXELS FOR VARIOUS STAGES**

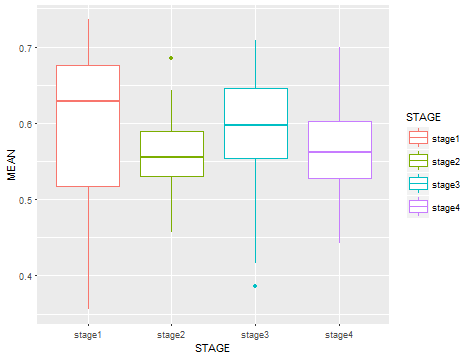


**BOX PLOT CONTAINING SSE FOR VARIOUS STAGES**

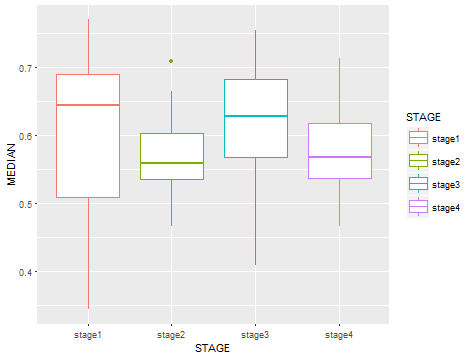


**BOX PLOT CONTAINING RADIUS FOR VARIOUS STAGES**

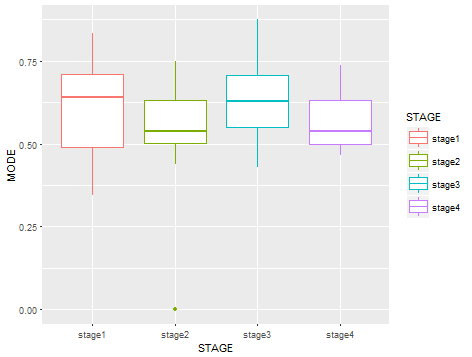
The above plot shows that stage1 values are relatively lower than other stages. For stage2, 50% of the values are lower than stage3 and stage4.However there seems to be an overlap of Stage2, Stage3 and Stage4 values.



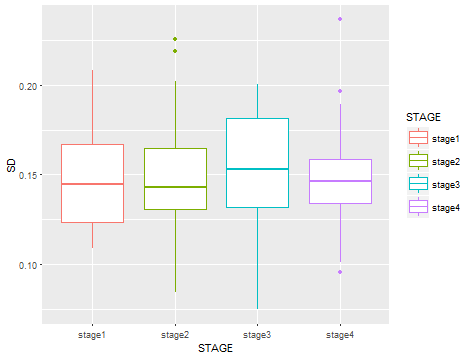
**BOX PLOT CONTAINING MEAN OF PIXEL INTENSITY VALUES FOR VARIOUS STAGES**



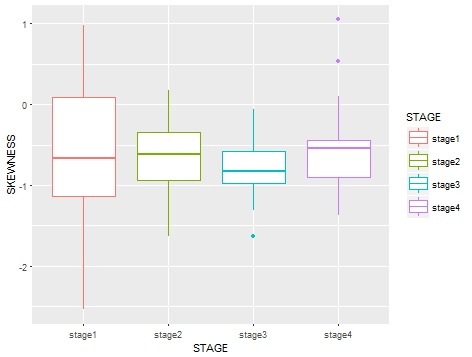
**BOX PLOT CONTAINING MEDIAN OF PIXEL INTENSITY VALUES FOR VARIOUS STAGES**



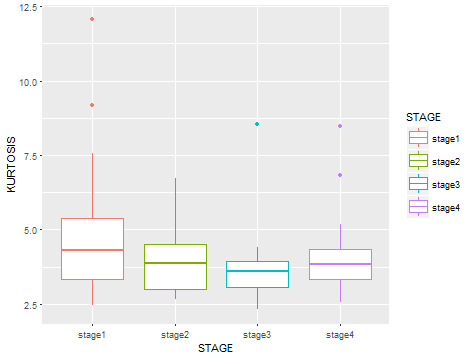
**BOX PLOT CONTAINING MODE OF PIXEL INTENSITY VALUES FOR VARIOUS STAGES**



**BOX PLOT CONTAINING STANDARD DEVIATIONOF INTENSITY VALUES FOR VARIOUS STAGES**



**BOX PLOT CONTAINING SKEWNESS OF INTENSITY VALUES FOR VARIOUS STAGES**



**BOX PLOT CONTAINING KURTOSIS OF INTENSITY VALUES FOR VARIOUS STAGES**

From the above plots it can be seen that only in case of Skewness there seems to be less intermixing between stage3 and stage4 cases. So this feature can be used to differentiate between the two.

**RESULTS**

MODELLING USING DECISION TREE

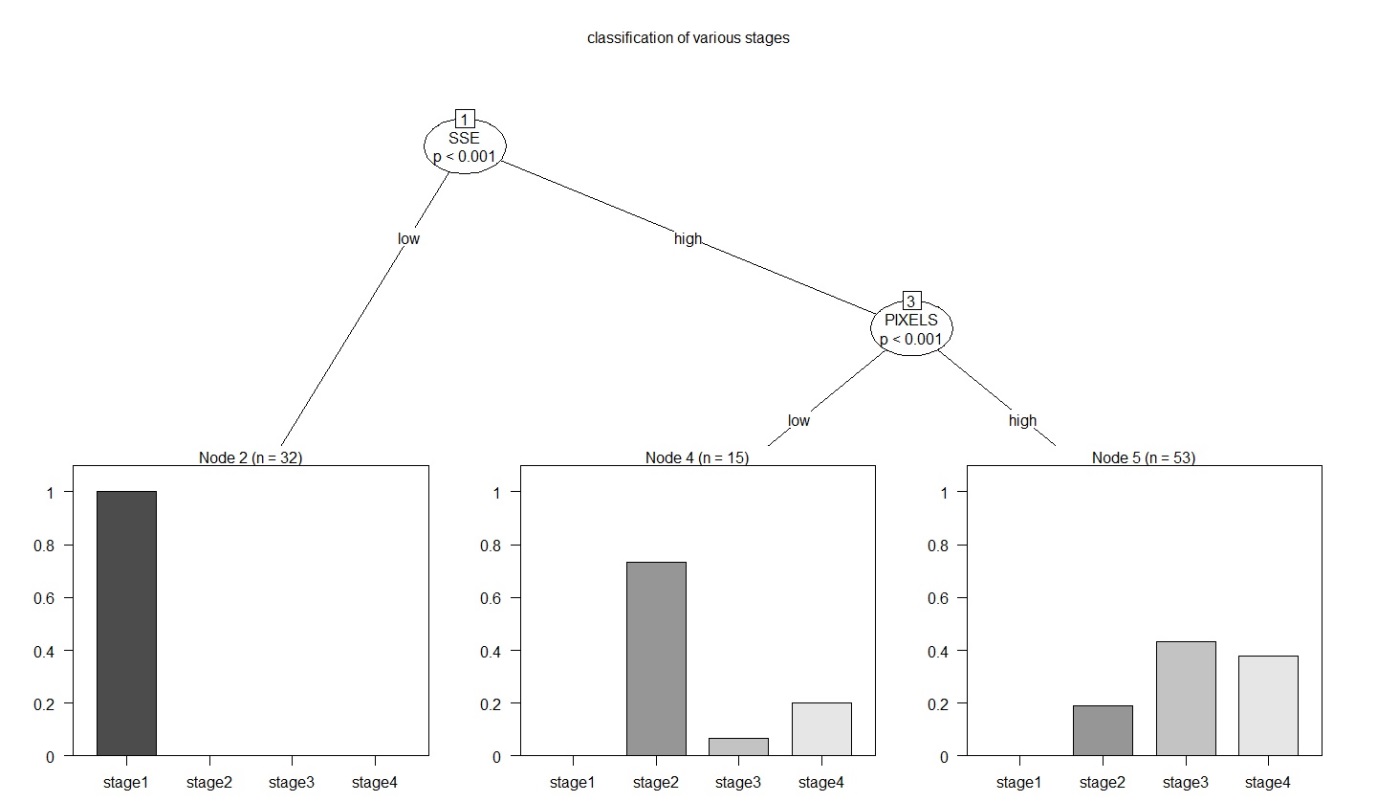
At first **Decision Tree** was applied on the size related features to separate stage1 and some of the stage2 cases. Before running the algorithm we need to convert the continuous variables to categorical data as it is a prerequisite in our model.

The following set of algorithms was used based on the box plots to convert them to categorical variables

* If(SSE)<max(stage1$SSE) then SSE=LOW else HIGH
* If(PIXELS)<median(stage2$PIXELS) then PIXELS= low else high
* If(RADIUS)<median(stage2$RADIUS) then RADIUS= low else high



THE ABOVE TABLE WAS USED TO BUILD THE DECISON TREE



PLOTTING THE DECISION TREE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **stage1** | **stage2** | **stage3** | **stage4** |
| **stage1** | **32** | **0** | **0** | **0** |
| **stage2** | **0** | **11** | **1** | **3** |
| **stage3** | **0** | **10** | **23** | **20** |
| **stage4** | **0** | **0** | **0** | **0** |

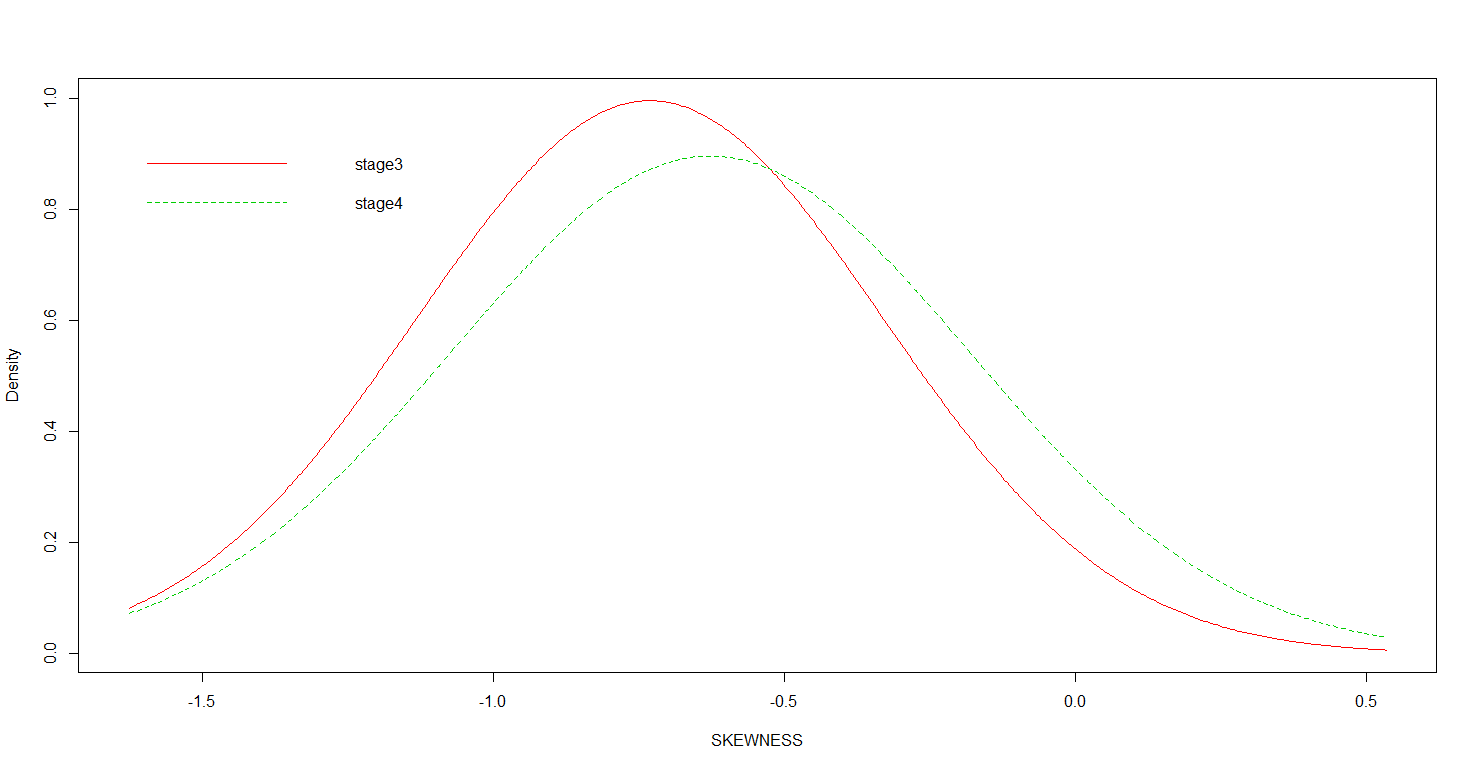
CONFUSION MATRIX

The above plot for decision tree and confusion matrix shows that we were completely able to identify the stage1 cases but in case of stage 2 where the size is comparatively lower than Stage 3 and Stage4 were identified with 73% accuracy.Stage3 and Stage4 cases were completely misclassified using the decision tree classifier.

The next step was to distinguish the stage3 from stage4 cases as decision tree was unable to do so. We used both **Naïve Bayes** and **KNN** classifier on the features related to statistical measures of pixel intensity values. The results are as given below-

MODELLING USING NAÏVE BAYES

There were 47 cases of stage3 (24) and stage4 (23) in the entire dataset. Out of 47 a total of 35 cases were taken using random sampling to train the model and the remaining 12 cases were used for validating the model. This step was performed again and again to find the highest, the lowest and the average accuracy of our model. The highest accuracy it achieved was close to **83.33%** and the lowest accuracy was around **41.67%**.On an average the accuracy was **60%**.Out of all the features “skewness” showed the maximum difference in mean between the two stages i.e. stage3 and stage4 as we had already seen from the box plots. Thus “skewness” was taken to measure the likelihood of both the classes.



**DENSITY CURVE FOR STAGE3 AND STAGE4 CASES**







**CONFUSION MATRIX FOR ACCURACY OF 83.33% AND 41.67%**

MODELLING USING KNN CLASSIFIER

The same steps that have been used in Naïve Bayes classifier has been repeated for modelling using KNN classifier. The highest accuracy was found to be around **91.67%** and lowest accuracy was found to be around **33.33%.** The optimal value of K is found to be **“7”**.







**CONFUSION MATRIX FOR ACCURACY OF 91.67% AND 33.33%**

COMPARING NAÏVE BAYES AND KNN CLASSIFIER

A comparative study was done to see which of the two performed better if the same training and testing sample was feed into the models. The results are as follows-

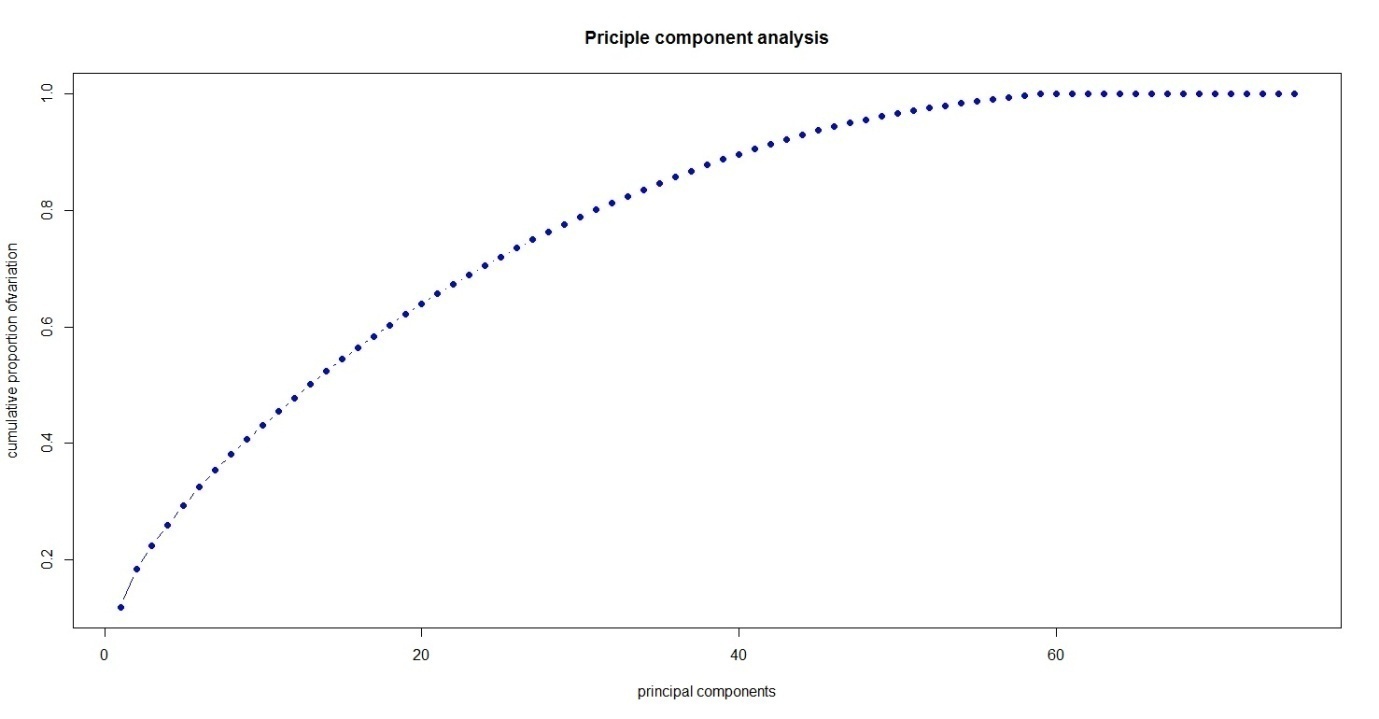


From the above table we can see that out of 10 different instances in 7 of them Naïve Bayes performed much better in classifying stage3 and stage4 cases than KNN. This shows the better predictive capability of Naïve Bayes compared to KNN.

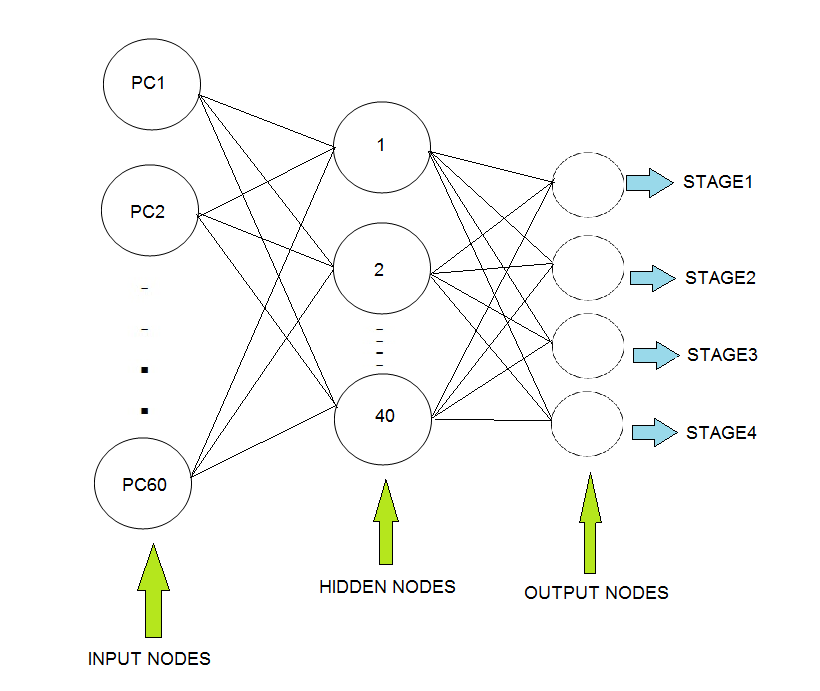
MODELLING USING ANN

In case of modelling using ANN we don’t need to extract features from the images or pixel values but instead the entire image or all the pixel values of the image can be taken as an input feature to feed into the network. The type of neural network which we would be using in our analysis is called **“feed-forward neural network with error back propagation** **algorithm”**. This kind of network like any other requires complex training procedure to find the optimal values of **learning rate**, **epoch** and **hidden nodes**. The procedure is to vary any one of them say epoch but keeping fix the learning rate and hidden nodes to find the optimal value where error is minimised .The same can be done for finding the optimal value of the other two. Theory suggests keeping the learning rate between (0.1-0.9) and hidden nodes less than the input features but more than the output nodes.

Another problem with ANN is that it is computationally very intensive, the more input features we add more complex the network becomes, and hence training the network becomes very difficult. In our study after standardizing the images to a fixed size of (250×250) pixels the input features came to be around “62500” which is very large and not possible to compute with our current system. So to reduce the dimension **Principle component Analysis** was done on the entire dataset (100 observations) from stage1 to sateg4 and the following results were obtained.



From the above plot we can clearly see that up to 60th principle component almost 100% of the deviation has been explained and beyond that the cumulative proportion of deviation is almost constant. So we take the first 60 principle components to feed into the neural network.



**FINAL ARCHITECTURE OF ANN MODEL**

The optimal number of hidden nodes is found to be “40” along with the learning rate set as “0.5” while the epoch is taken as “15”.Logit function or Sigmoid function is used as the activation function in our model.

The entire dataset has been split into two groups one for training the network and the other for validating our model. The training dataset comprises of 75 observations selected randomly from the entire dataset while the validation set contained 25 observations different from the training samples. The results are given below.

|  |  |
| --- | --- |
| **original** | **ANN classifier** |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage2 | stage3 |
| stage2 | stage3 |
| stage2 | stage3 |
| stage2 | stage3 |
| stage2 | stage3 |
| stage3 | stage3 |
| stage3 | stage3 |
| stage3 | stage3 |
| stage3 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |
| stage4 | stage3 |

Accuracy=52%

The reason for this is that in the training sample the weight age of Stage1 and Stage3 cases were more, so the weights in the network have been adjusted to classify them more accurately than stage2 and stage4.The table below shows the various instances in the training sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **stages** | **stage1** | **stage2** | **stage3** | **stage4** |
| **Training instances** | 23 | 16 | 20 | 16 |

Let’s try another instance where the training sample weightage is different as given below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **stages** | **stage1** | **stage2** | **stage3** | **stage4** |
| **Training instances** | 24 | 15 | 17 | 19 |

The results after training the model is given below

|  |  |
| --- | --- |
| **original** | **ANN classifier** |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage1 | stage1 |
| stage2 | stage4 |
| stage2 | stage4 |
| stage2 | stage4 |
| stage2 | stage4 |
| stage2 | stage4 |
| stage2 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage3 | stage4 |
| stage4 | stage4 |
| stage4 | stage4 |
| stage4 | stage4 |
| stage4 | stage4 |

Accuracy=48%

In this case since the weight age of stage1 and stage4 is more the weights in the network have been adjusted to classify them more accurately than stage2 and stage3.

**SUMMARY OF THE MODEL**

Accuracy-52.4%

Accuracy-100%

Accuracy-100%

**CONCLUSION**

* Stage1 cases have been completely identified both using Decision tree and Neural networks with **100%**Accuracy.
* Stage2 cases where the size is presumably lower than stage3 and Stage4 cases have been identified using Decision tree with **73%** Accuracy.
* For separately classifying the Stage3 and Stage4 cases Naïve Bayes classifier has been found to be more accurate than KNN. This shows that in case of intermixing Naïve Bayes is more adequate to use.
* ANN model depends a lot on the training dataset. The accuracy of the model increases with increase in training instances.
* ANN model is very hard to train and computationally very intensive than other models but with more training instances the predictive capability of the model increases.

**IMPLICATIONS**

The study indicates that using M/L techniques in diagnosing renal cell carcinoma cases has been found satisfactory. Even though the dataset contained very few observations the results were found to be adequate. However the accuracy could be improved with more sophisticated techniques like “deep learning” and availability of more data.

**REFERENCES**

1. KidneyRenalclearcellcarcinoma,https://wiki.cancerimagingarchive.net/display/Public/TCGA-KIRC/
2. HowToImplementNaiveBayesFromScratchinR,<http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>
3. Introduction to K-nearest neighbours simplified, <https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/>
4. Tariq Rashid, “Make your own Neural Network,”, Ed. 1st, New York: Wiley, 1997, pp. 1-222.
5. Toby Segaran, ”Collective intelligence”, Ed. 2nd, O’REILLY, pp. 1-323.
6. Basic operations on images, <http://docs.opencv.org/3.0beta/doc/py_tutorials/py_core/py_basic_ops/py_basic_ops.html#accessingimageproperties>.
7. Basicimageoperationspixelaccess,<http://www.bogotobogo.com/python/OpenCV_Python/python_opencv3_basic_image_operations_pixel_access_image_load.php>.
8. Diagnostic Image analysis group, ”A survey on Deep learning in Medical image analysis” Radbourg university Medical centre, pp. 1-5.
9. Wes Mckinney, ”Python for data analysis”, Ed. 2nd ,O’REILLY, pp. 45-78.
10. Garrett Golemund, ”Hands on programming with R”,Ed. 2nd ,O’REILLY, pp. 4-23