FINAL REPORT

DAB402: CAPSTONE PROJECT

SUBMITTED BY

KOMALPREET KAUR (0733909)
SEOUNG KYOUNG RYU (0725164)
GURTEJ SINGH UBHI (0734821)

APRIL 17, 2020

CONTENTS

INTRODUCTION

Health care is one of the most widespread domains that is gaining attention these days. It cannot be denied the fact that heath care plays critical role in the improved principles like diagnosis, treatment and maintenance. An efficient health system is highly important for a nation economy. With the advancement in the information technology , it becomes quite easy to equip it with the health care and work for the benefit of the mankind. At the same time the role of data science in the field of health care cannot be ignored. It is bringing a revolution by collaborating solutions based promising resources and implementing various alternatives to a new level. Healthcare analytics is a new field emerging that will not only be providing helpful solutions which can be customized accordingly to the needs.

Moreover, the personalised solutions that are provided by the healthcare analytics help one to understand the actual importance in day to day run activities. This not only cover the variety of diseases but also provide recommendations based upon the likelihood of their occurrence. In the health care there are couple of diseases that have huge impact on the livelihood of people and their ramifications are often eluded. Say for instance the death rates due to cancer is on rise and there exist so many kinds of cancer which are prevailing, and their critical impacts are really astonishing.

In this capstone project , we have tried to work upon the breast cancer. Since in the modern world ,this kind of cancer has a ravishing impact on the lives of people. The impact and its outcomes cannot be outlooked . We will be working on the various reasons for causes for which the breast cancer occurs. The application of the machine learning will take place with the equipment of the big data analytics that will yield better results for the betterment for the future outcomes and help in generating the alternatives that are supportive in nature will also coordinate for the betterment of the society.

To add upon , we have variety of machine learning models that are available. It becomes sometimes tough to understand which algorithm is applicable. The applicability of the dataset depends upon the nature of the dataset and the amount of the precision needed to carry out the process of the modelling. Also, at the same time it is quite important to understand the nature of the dataset. The dataset plays a crucial role in the carryout of the process of predication for prognosis and diagnosis. For the variety of algorithms that are available, like in the case of supervised learning ,

classification and regression have been widely used for the processing of results. And in the case of the unsupervised learning , clustering is one of the most popular algorithms that is used in that case.

In this capstone project , we are going to use classification models and going to employ various accuracy metrics to carry out the comparison process and make the prognosis process go smoothly. At the same time, we are going to work on various ways by which the process of optimisation can be achieved. The various way by which the comparison can be done will the matter of outcome that will be helpful in generating the results.

For the decision-making process , all the summary and the comparative results will be used to carry out the better outcomes that will be used for the generalisation of the idea about the breast cancer. Such a kind of the validations will be helpful in the cure and treatment for the breast cancer and provide a kind of the framework for the utilisation by the people and will be helpful to set out the recommendation in such a way that an non-IT person can use it for the better understanding of the knowledge about the breast cancer and will take appropriate measure to tackle it. Such a kind of algorithm will set up an example that how the data analytics and health care can be leveraged to find out better findings to understand the health care industry and take better decision that are not only optimised in nature but also less complex.

DESCRIPTION ABOUT THE DATASET

In the current problem , the dataset has been collected from the Kaggle[1] , the dataset has 570 rows and 32 columns. For the dataset , the features have been computed from a digital picture using the FNA (fine needle aspirate) methodology of the given breast mass. Talking about the FNA , it is a medical procedure in which a thin needle is inserted into the area of the body with has irregular seeming mass or the tissue, this method is employed to look for swelling. This dataset has been collected with the help of this methodology. The 32 attributes are defined as

ID Number: This is the id of the patient as is utilised as a primary key .

Diagnosis : This will bifurcate as Malignant (which means the tissue is composed of the cancer cells and it can affect the nearby tissues as well) and other one as Benign (which means it will not spread nearby).

Further to this , there are values that are defined as:

Radius : This is the calculated mean of the distances from the centre to the set of the points on the perimeter.

Texture : This is the calculated standard deviation of the grey-scale values.

Perimeter: This is the statistical measure of the tissue .

Area: This is also a statistical measure that is calculated based upon the points of the tissue.

Smoothness: This is defined as slight dissimilarity in the length of radius.

Compactness: This is also a statistical measure that is calculated based upon the points of the tissue.

Concavity : This is the sternness of the points that are hollo-shaped in nature.

Concave Points: This is the collective points of hollow-shaped portions.

Symmetry: If there is any mirror image formation , if it forms any mirror image , they are symmetry in nature else they are not.

Fractional Dimension: This is defined as the measure of the figures depending upon their complexity , which is actually degree of count of points that lie on a given point.

- More significantly , all the features values are calculated up to 4 precision. points.
- Also , mean , standard error and worst of these structures have been calculated for each image.

Talking about the ethics(5C's : Consent , Clarity, Consequences ,Control , Consistency) in data collection. As it is clear that the data is collected from Kaggle, it is open so it will not hamper the Control of ethics. Consent part has also verified since the data is free from bias and equal consent has been taken that the data will be used for the purpose of research only. Consequences , as far as this is concerned, the results are clear and there will no manipulation in the data and the results will be carried out for the purpose of finding insights only. Clarity has been explained clearly since the data is transparent in nature this will not hamper the confidentiality concerns. Consistency , on the other hand with each result , they will be similar in the nature and this will not deviate the results in any cost. All the aspects of ethics have been taken into this during this capstone project and it has been made sure that that whatsoever the results will come , it will be helpful for the betterment of the analysis that has been carried out previously.

RELATED WORK

Following section will elaborate the related work in this field .

PAPER-I

Instead of using a full biopsy[2], Fine needle aspirations (FNAs) enable to inspect a minor quantity of tissue from the tumor. For the analysis, they used DIP and ML methods. First step was the finalizing frontier of cell nucleus by a dynamic model.

The initial step is to draw approximate boundary which is rough initial outline. After the initial step, it finishes drawing boundaries. The interactive process took 2 ~ 5 minutes.

They use following formula to generate boundaries.

$$E = \int_s (\alpha E_{cont}(s) + \beta E_{curv}(s) + \gamma E_{image}(s)) ds$$

To determine the boundaries, they use 3 elements, which are continuity, curvature, Image. The computation will be conducted till all arguments localise for a least value of energy. They extracted features from the image, which are radius, perimeter, area, density, and other parameters. Finally, the classification procedure used is a variant on the Multisurface Method (MSM)10' known as MSM-Tree 23.

The technique uses a linear programming model for separating planes. In this classifier, predicted correctness was 97%.

For the linear classification, they used 2 factors, sensitivity, specificity to identify. They have applied this machine learning techniques to Wisconsin Hospital to determine and diagnose breast cancer.

PAPER-II

In this research paper, the author[3] used the image-based dataset to implement the methodology to diagnose the breast cancer. In addition to this, he employed various mathematical models to test the diagnostic system. Just to create the image dataset , fine sample was taken in the form of fluid as the measuring sample. Then for the purpose of visualisation, sample nucleus was seen under the camera for the purpose of data identification.

The parameters were measured with the shape and size. Based upon the used model that are mathematical in nature to carry out the analysis. Later on, a classification method was used to classify among the various repeating and non-repeating features. The idea behind was the linear programming that used plane method to divide it into two halves with the equidistant method.

The precision value is measured based upon the method of cross validation which used all the folds that were created using the aforementioned method. However, when these samples were visualised under the microscope, a thin grey line was seen that meant that the data was fine tuned and much more in granule level. The argument of parameter was much more sophisticated and had more detailed features for forecast and analysis. At the end, the results yield much more better results as they had more values of correlation. With the increasing values of parameters, it becomes crystal clear that such a kind of study becomes helpful in understanding the various hidden features that are fruitful in understanding the diagnosis strategy for breast cancer.

PAPER-III

In this research paper[4], computer system was made to diagnose the breast cancer. It was made with a combination of ML increase the accuracy of the breast cancer based upon various strategies that are defined to understand the various processes in depth.

In the initial stages, fine image of the data was taken to elaborate and understand the process in detail and the feature extraction was done based upon the radius which is measured by the distance of the centroids. Later on, the area was calculated on the adjusted parameters the compactness was calculated to ensure that they leverage well on the centroids of the parameters. The classification procedure was later used on the basis of the tree that made decisions as per the distribution curve by projecting the points that shared the same dimensions and the estimated dimension of malignancy was calculated on the planes that used to divide the surface area above the region in a better way accordingly to the zones of the places.

In the later stage, these so-called collaborated points were made to establish a link among the various target that were set, and the measure accuracy was calculated. The proposed system was implemented on a normal UNIX based system and the elements were used to predict the accuracy and elements in the necessary format to allow the users to use them for their ease and benefit. Also, such a kind of system will help the better forecast and prediction in the era of technology when it was quite tough to predict such kind of disease in the health care industry.

PAPER-IV

In this research paper [5] , the author worked for increasing the accuracy of the images that are based on the FNA methodology. They used these images to convert these into a useful feature and build models. Later on, a collaborative effort was used with an additional machine learning models to calculate these accuracies in well detail manner. Data analysis was done with the help of the machine learning models like they employed linear programming models in respect to the data gathered along with the features used.

In the earlier phase, 3D structure was used to gather the data in a multidimensional structure such as to gather those pigments into a better way. After that these features were seen under the microscope, the simpler and complex system was built to analyse and find the hidden traits along with the similarities. The mechanism used was much more similar to finding the contextual features that are much more elastic in nature.

In the final stage , the combinational probability was calculated on the entire dataset using the mathematical fundamentals of likelihood of the centroids along the radius of the portion of the datapoints being calculated for the purpose of the smaller distances. This probability of estimation was used to evoke the mathematical notations in much more appropriate manner such that there exist the variations in the splitting of the marginal datasets for the user defined locations. The consistency doubled along with the efficiency of the various parameters that were used to fine tune the datapoints in the near future. It is obvious to understand that the nature of the dataset becomes important part to understand the various reasons of study of the feature selection.

PAPER-V

In this given article , the author [6] has used computer-based approach to devise strategies that are helpful in classifying the cancer from non harmful cancer cells. Various approaches have been used to correctly segregate the features in the best possible manner. In the paper the author used to classify the aspirates in the best possible manner to help to evaluate the model learning process. The historical evidence has made it easy to understand the features selection process in much detail process.

The simpler classifier process made it easy to cross validate them better. They used the logistic approach to make sure that the parameters should work in a better way accordingly to ensure the better development of prediction. Feature analysis is one of the critical components that takes place to make sure that these diagnostic features work in a well-defined manner. So, the features were divided to make sure that they work in the plane division works better along with the options available to diagnose better.

Also, at the same time , these features worked better in a way so as to reduce the problems of lower accuracy and predictive analysis. The values of correlation when coupled with the accuracies so that the predictive analysis work in a much more accurate way. Selection of the nucleus in the aperture becomes more sufficient and the probability of the approximation becomes appropriate as these values are nearby the outliers so that these may be used for the better decision and prediction of the breast cancer for the prognosis and detection for the betterment of the health care industry.

PAPER-VI

In this research paper, the author [7] used machine learning methods to predict the breast cancer. The EM (Expectation Maximisation) calculation is a strategy for effective estimation from deficient information. In any inadequate dataset, there is aberrant proof about the presumable estimations of the surreptitiously values.

This proof, at the point when joined with certain presumptions, includes a prescient likelihood dispersion for the missing qualities that ought to be arrived at the midpoint of in the factual investigation. Various data mining methods have been used to analyse which model works better and why depending on severity and complexity of the dataset. Each method had it own risks due to which they are applied actually at the area of interest. Also, in such a comparative way the nature of algorithm may vary the need and desire to understand various issues that are related to the complexity of dataset. The performance of SVM is high accurate in all the cases and this is due to the fact that even due to the presence of worthless in the dataset and the outliers, there were some aspects that made it easy to predict and predict the features in a better way. This accuracy can be improved with the combination of back propagation and feature elimination methods that are widely sink with the dataset. All these results were achieved using the help of 10-fold cross-validation for the calculation of impartial accuracy that is generated during the prediction of accuracy of each model during the stage of evaluation and processing.

PAPER-VII

In this research article, the author [8] used a combinational approach for detection of breast cancer based on prediction and diagnosis. They mainly focussed on classification since it forms the basis of the data mining methodology. This was computed on WEKA – which is a collection of various machine learning methods that are employed for data preprocessing and cleaning.

In order to study the actual impact of the error on the dataset, various measures were employed to detect the relation with these factors in depth. To analyze the efficiency, ROC curve was used to employ the need to detect any deviation from the performance of various other classifiers that produce better results. Thus, this was significant important as the comparison becomes much clearer based upon features and patterns if they are visually represented in the form of figures and diagrams.

The experimental results were carried out in a well-defined framework and clear images were used to figure out the need of various parameters interested to follow up for the values for efficiency and effectiveness. SVM (Support Vector Machine) performed the best in all the cases due to higher values of recall and f-scores as predicted from the confusion matrix.

Also, low error rate has been a significant role in playing performance for the parameters of accuracy, precision in the process of classification of the datasets in a better way. The classes that were predicted better for Malignant and Benign had the higher accuracy in feature selection and other important parameters that are quite useful in detecting the prediction and diagnosis.

PAPER-VIII

This research paper [9] is a combination of various methods of machine learning to improve the accuracies that are significantly important for Breast cancer diagnosis. In the first section of the paper, machine learning algorithms have been used to significantly improve the accuracy and precision. This worked with the combination of various ML algorithms based on nature and type of dataset that was used to enquire the study of Breast cancer. In the later part of research paper, it focussed on the working of neural networks and learning rules. This was much similar to any mathematical analysis which means that the weight and other parameters can be adjusted on the basis of need and desire of performance of statistical models.

In the middle section of the paper, other complicated models were employed to make sure that they fit well in each and every phase of data modelling and feature selection. Also, filters were used to employ that they work well even in digital images in the form of complex patterns where it becomes tough to understand the nature of dataset.

In the last section of the research paper, vector machines were employed to understand the impact on regression and classification. Based upon all the comparative study and analysis it can be concluded that SVM (Support Vector Machines) was the best option to choose from other options due to accurate calculations due to optimised calculations and fast applications of feature selection methodologies due to low response time and high parameters of interest of error (they yield low values of error and high precision rates of calculations)

PAPER-IX

In this research article author [10] developed a system which is automated and is a combination of the rules and ML methods. This was quite novel way of using the features in such a way this will reduce the chances of lower accuracy issues and make best decisions. This process was divided into various stages. In the first one, the data was reduced to lower scale to have a better idea on what the data is all about. This helped the complete process to reduce the training time to manifolds. Even if we have higher number of training data , the accuracy will not reduce, rather it will help to penetrate more by reducing the efforts used in selecting the feature data. After that the Bayesian Probability was used to get the ratios in detail for the various feature selection and feature impurity. Then a method of ranking was used to get the data in much details and to rank them accordingly.

After than at the later stages, feature elimination was used to understand the contribution of each feature in detail so that the accuracy could be figured out better. So, in the last stage it was concluded that such a kind of automated system will be an optimum tool for many of the hospitals in understanding various issues. This was made possible only with the help of the combination of various features in response to accuracy for prognosis and diagnosis.

The expert system that was created was legacy system that was a best to used only with the viability of predefined features that are available in the dataset. They made accurate decisions also to improve the accuracy of the medical data that could be examined in shorter time with utmost accurate features and better availability of the domain knowledge.

It provided not only better results but also accurate calculation at the shorter range of time even for the complex data.

PAPER-X

In this research paper [11], the author tried to crate a system which was based upon the combination of ML and a rule. This was necessary as there was much need to develop a system that should be reliable enough to predict the things accurately. This was based upon the consideration of studying deep about the patterns and the relationships of the history of the patients that was kept as a record.

Also, at the same time, plethora methods were used with algorithms to make sure all the parameters of evaluation yield better results. So, at the end the model that performs best and gives the best results will be selected as the best model for prediction. Cross-Validation was also used to make sure that the methods that have been employed works well in all the possible cases of model accuracy and reliability. After all the analysis was done , based upon the values available it was proved that TRF model worked well in all the cases due to some factors. First and foremost, this method was very fast to used and it used less compassion in finding the features. The data that has been provided as input do not need any special type of scaling, the data that is available is already preprocessed so no need of scaling is required for the later stages.

For the purpose of memory features and implementation, the storage constrains are less in number, so they are more efficient. So, it can be said that TRF (Tree Random Forest) model works better for the Breast cancer survival prediction. Various hospitals have used this methodology to implement this model to understand various factors in detail. For the precision and better evaluation of the cases, the clinical data can be used with the combination of other data as well for the purpose of better decision-making purposes.

METHODS

The first and the foremost thing that was done is EDA (exploratory data analysis). This is necessary to understand the nature of dataset as it helps in finding the hidden trends and relationships in the dataset. After that data visualization and model processing.

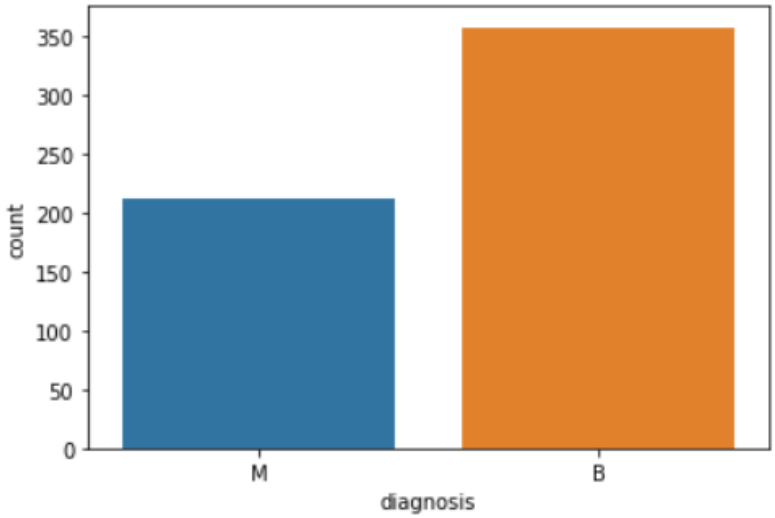

Fig : Plot for the count of Malignant and Benign Cancer

From the above graph it is clear that Benign cancer has more count when compared to Malignant. This will help to look for features that will more towards mitigating the causes of cancer that is malignant in nature.

Heat maps (also called correlation graph) are also very important in understanding the correlation between various features that are present in the dataset. Once the heat map is created it will highlight the features that play an important role in the feature selection.

- Here correlation graphs are used so that we can remove multi collinearity (i.e. the columns that depend on each other) because using same columns twice is not useful.
- The following heat map depicts the relationship between various features on the basis of mean.
- The diagonal has the value 1 (dark color) and each variable shows strong correlation to itself.
- On the other hand, higher the number and darker the color shows higher correlation between the variables.

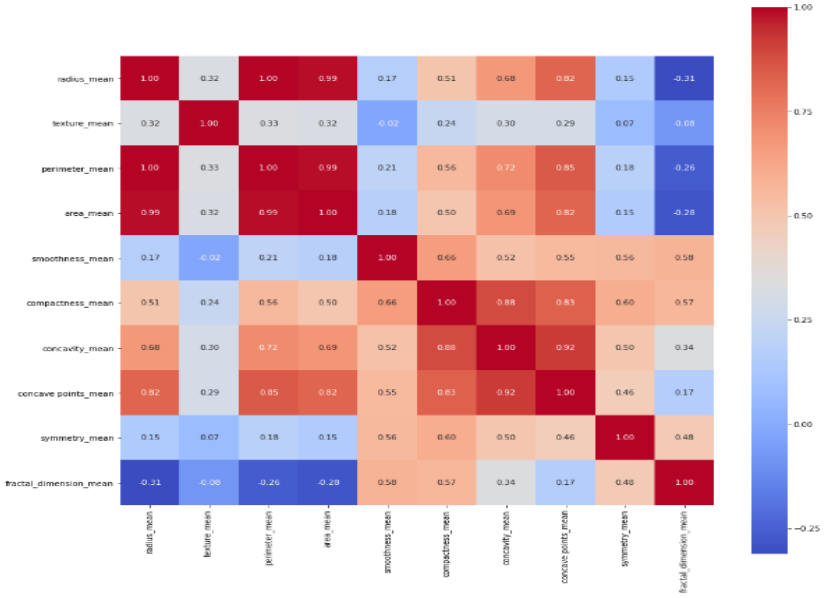

Fig : The heat map depicts the relationship between various features on the basis of mean

From the above graph, we can see that the mean features show strong correlation as most of the values are higher and falls between the range of 0.5 to 1.

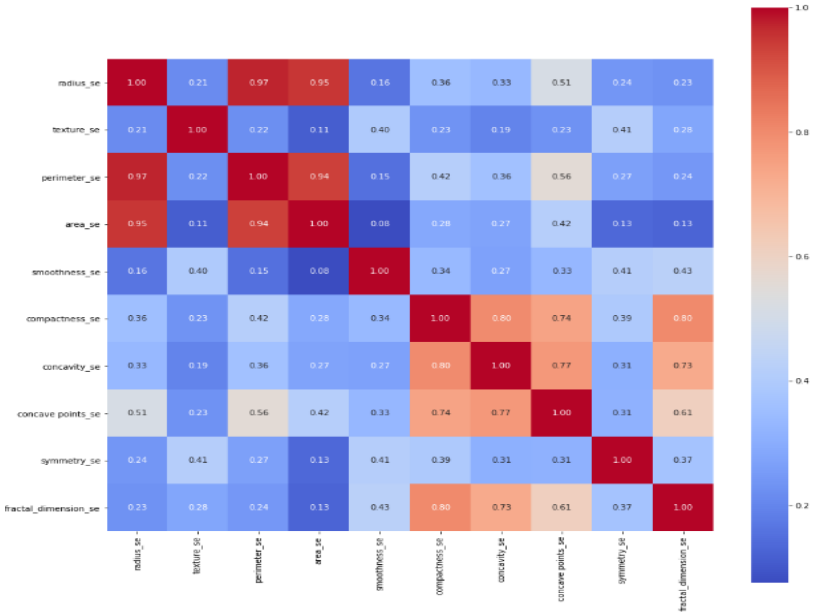

Fig : The heat map depicts the relationship between various features on the basis of standard error.

From the above graph, it can be depicted that the standard errors have less correlation between the features as most of the values fall under the range of 0 to 0.6.

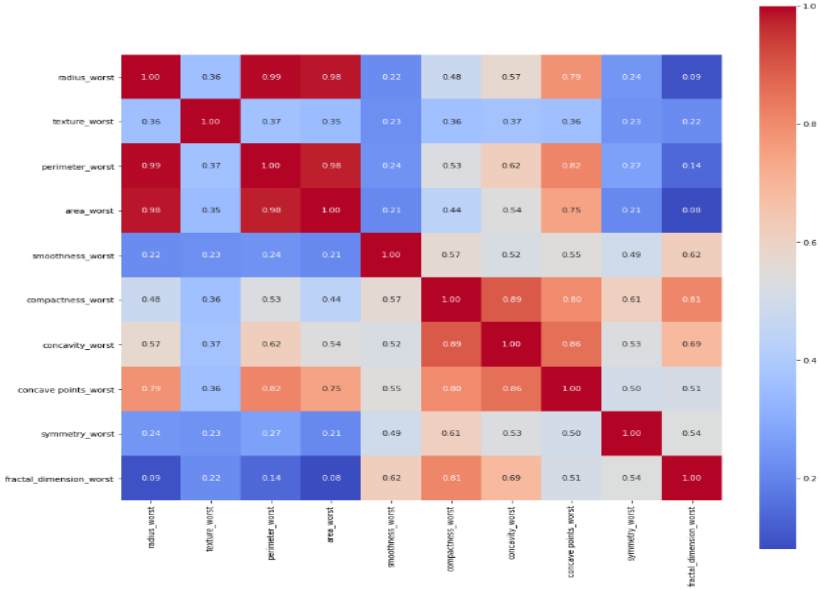

Fig : The heat map depicts the relationship between various features on the basis of worst features.

From the above graph it shows that the features have less correlation in the dataset as most of the values fall under the range of 0 to 0.6.

From above graphs, it is quite clear that mean features shows high correlation as compared to standard error and worst features.

Following set of features shows strong correlation with each other so we have used one feature instead of three.

- radius_mean, perimeter _mean and area _mean
- compactness_mean, concavity_mean and concave points _mean

Therefore, following features are selected to train our models:

- 'radius_mean', 'texture_mean', 'compactness_mean', 'symmetry_mean', smoothness_mean'.

```
from sklearn.model_selection import train_test_split

train, test = train_test_split(cancer, test_size=0.25)

print("Training set shape:{}".format(train.shape))
print("Test set shape:{}".format(test.shape))
```

```
Training set shape:(426, 31)
Test set shape:(143, 31)
```

```
prediction = ['radius_mean','texture_mean', 'compactness_mean', 'symmetry_mean', 'smoothness_mean']
```

```
x_train = train[prediction]
y_train = train.diagnosis

x_test = test[prediction]
y_test = test.diagnosis
```

Here the prediction array consists of the features that are selected from the heat map.
The data is divided into training set and test set in the ratio of 75:25 respectively.

MODEL IMPLEMENTATION

In this section basis , model implementation will be covered and after that comparison will be done to carry out analysis. In this we have implemented 5 types of models :

- Random Forest Classifier
- Logistic Regression
- KNN
- Decision Tree
- Neural Network (MLP)

RANDOM FOREST CLASSIFIER

This is a supervised machine learning algorithm that is based on various decision trees from a randomly selected subset of the training set. This further uses averaging method to improve prediction power and accuracy.

The accuracy of the Training set is : 99.8%

The accuracy of the Test set is : 93.0%

Model Evaluation

a) Classification Report

	precision	recall	f1-score	support
B	0.93	0.97	0.95	96
M	0.93	0.85	0.89	47
micro avg	0.93	0.93	0.93	143
macro avg	0.93	0.91	0.92	143
weighted avg	0.93	0.93	0.93	143

Fig : Results for various scores accuracy

Precision is the percentage of the result that are relevant . Recall is the percentage of the total relevant results correctly classified by the algorithm. and if the F scores reaches best value i.e. 1 , it means perfect precision and recall.

For instance , from the above classification report it can be concluded that precision here is 93% which means our model will have 93% relevant results whereas , 97% of recall means it will generate 97% relevant results out of the total classified values in the case of Benign (B). Same is the case with Malignant(M).

Also, after running the confusion matrix code , following output was produced.

a) <u>Confusion Matrix</u>

```
[[93  3]
 [ 7 40]]
```

Fig : Confusion matrix for Random forest classifier

From the above matrix we can identify that our model performs well since the number of correct predictions(93,40) is greater than the number of incorrect prediction(3,7).

LOGISTIC REGRESSION

This is a statistical machine learning model that employs the usage of probability and predictive analysis. This uses a linear model to find the relationship between dependent and independent variable. This can be extended to several classes as well.

The accuracy of the Training set is : 89.2%

The accuracy of the Test set is : 90.9%

Model Evaluation

a) **Classification Report**

	precision	recall	f1-score	support
B	0.90	0.98	0.94	96
M	0.95	0.77	0.85	47
micro avg	0.91	0.91	0.91	143
macro avg	0.92	0.87	0.89	143
weighted avg	0.91	0.91	0.91	143

Fig : Results of accuracy

From the above classification report it can be concluded that precision here is 90% which means our model will have 90% relevant results whereas , 98% of recall means it will generate 98% relevant results out of the total classified values in the case of Benign (B). Same is the case with Malignant(M).

b) **Confusion Matrix**

```
[[94  2]
 [11 36]]
```

Fig : Confusion matrix for Random forest classifier

From the above matrix we can identify that our model performs well since the number of correct predictions(94,36) is greater than the number of incorrect prediction(2,11).

K Nearest Neighbor (KNN)

This is the third model that has been used. KNN is a kind of algorithm that don't depend upon the parameters. This model is said to be lazy because it takes time in understanding the patterns. This is used in the machine learning to predict the classification of new data point in the dataset. It splits the dataset in such a way to reduce Euclidean distance between each data point and the associated central measure of its cluster.

The accuracy of the Training set is : 90.8%

The accuracy of the Test set is : 90.2%

Model Evaluation

a) **Classification Report**

Following is the description of various scores calculated for logistic regression

	precision	recall	f1-score	support
B	0.91	0.95	0.93	96
M	0.88	0.81	0.84	47
micro avg	0.90	0.90	0.90	143
macro avg	0.90	0.88	0.89	143
weighted avg	0.90	0.90	0.90	143

Fig : Results for various scores

From the above classification report it can be concluded that precision here is 91% which means our model will have 91% relevant results whereas , 95% of recall means it will generate 95% relevant results out of the total classified values in the case of Benign (B). Same is the case with Malignant(M).

a) **Confusion Matrix**

```
[[91  5]
 [ 9 38]]
```

Fig : Confusion matrix for Logistic Regression

From the above matrix we can identify that our model performs well since the number of correct predictions(91,38) is greater than the number of incorrect prediction(5,9).

DECISION TREE

A decision tree is a supervised machine learning model that is build on if else conditional statements. It breaks down the data into smaller subsets and at the same instance a decision tree gets generated. The best part of decision tree is visualisation of tree like structure that makes more optimised results and better predictions.

The accuracy of the Training set is : 97.2%

The accuracy of the Test set is : 90.9%

Model Evaluation

a) Classification Report

	precision	recall	f1-score	support
B	0.93	0.95	0.94	96
M	0.89	0.85	0.87	47
micro avg	0.92	0.92	0.92	143
macro avg	0.91	0.90	0.90	143
weighted avg	0.92	0.92	0.92	143

Fig : Results for various scores

From the above classification report it can be concluded that precision here is 93% which means our model will have 93% relevant results whereas , 95% of recall means it will generate 95% relevant results out of the total classified values in the case of Benign (B). Same is the case with Malignant(M).

b) Confusion Matrix

```
[[91  5]
 [ 7 40]]
```

Fig : Confusion matrix for Decision Tree

From the above matrix we can identify that our model performs well since the number of correct predictions(91,40) is greater than the number of incorrect prediction(5,7).

MULTI LAYER PERCEPTRON (MLP)

Multi Layer Perceptron belongs to the class of feedforward artificial neural network. This makes use of supervised learning model which is known as back propagation for training. It helps in distinguishing the data that is not linearly separable . this is made up of three layers which is input layer , hidden layer and output layer.

Once the data has been split into training and test sets. There are two key points that have been added:

1) Sigmoid function is an activation function that produces output between 0 and 1 and is used on output layer. This is used for the models where we can predict the probability as the output.

```
def sigmoid(z):
    s = 1/(1+np.exp(-z))
    return s
```

Fig : Snippet of Code for the sigmoid function

The sigmoid function is also known as squashing function. So, it basically squashes the value between 0 to 1.

2) We have used tanh() function on hidden layer instead of sigmoid function because it gives more centred values. This is rescaled sigmoid function used to get more precise probabilistic values.

```
def tanh(z):
    s = (np.exp(z) - np.exp(-z)) / (np.exp(z) + np.exp(-z))
    return s
```

Fig : Snippet of Code for tanh function

Once the activation function is defined it is the time to define the forward propagation algorithms

In order to improve our code, we will be using a vectorized implementation of forward propagation. To do so, the following formulas will be applicable:

x is defined for the purpose of matrix multiplication

Layer 1

Z1 = W1.T x X where W1 signifies the matrix of weights in L1, and X signifies the feature matrix of measures

A1 = tanh(Z1 + b1) where b1 characterizes our intercept term for the first layer

Layer 2

Z2 = W2.T x A1 where W2 represents the matrix of weights in L2

A2 = tanh(Z2 + b2) where b2 characterizes the intercept term for the 2nd layer

Layer 3

Z3 = W3.T x A2 where W3 represents the matrix of weights in L3

A3= tanh(Z3 + b3) where b3 represents the intercept term for the 2nd layer.

```
def forward_prop(X,W1,W2,W3,b1,b2,b3):

    #First layer forward propogation
    Z1 = np.dot(W1,X)
    A1 = tanh(Z1 + b1)
    #Second layer forward propogation
    Z2 = np.dot(W2,A1)
    A2 = tanh(Z2 + b2)
    #Third layer forward propogation
    Z3 = np.dot(W3,A2)
    A3 = sigmoid(Z3 + b3) #A3 will produce our probability vector

    cache = {
                    "Z1": Z1,
                    "A1": A1,
                    "Z2": Z2,
                    "A2": A2,
                    "Z3": Z3,
                    "A3": A3
            }
    return cache
```

Fig : Snippet of Code for forward propogation function

With the forward propogation function , it keeps the calculation as well as the storage of intermediate values. Now that we have our forward propogation algorithm, we can proceed with back propogation and gradient descent. Succeeding steps will be used to perform gradient descent, and determine the improved weight values for every layer:

1) Starting by arbitrarily resetting our weight and intercept parameters
2) Running forward propogation through our neural network
3) Computing the derivatives of our weights and intercept parameters via back propogation
4) Improving parameters using derivatives from (3rd point).
5) Repeating 1 - 4 x times.

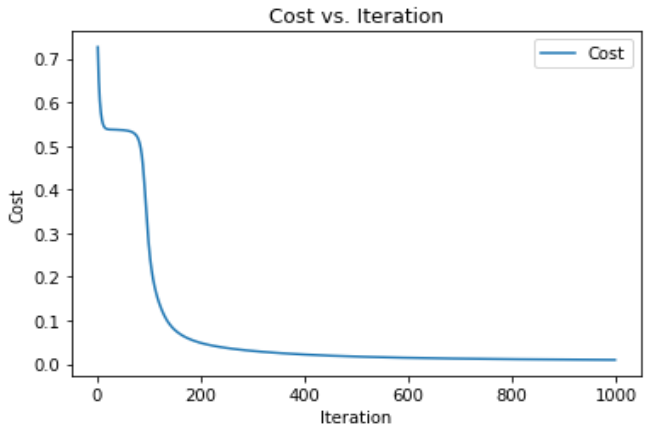

Fig : cost v/s iteration

- To check the working of algorithm, we should see a declining learning curve with iteration, which eventually flattens out. So, the above graph declares that the algorithm we have created is working as expected.
- This will help us determine a suitable number of iterations to run to determine the appropriate parameters. Once the code gets executed , following outputs is received.
  The accuracy of the Training set is : 98.95%

  The accuracy of the Test set is : 96.81%

```
164 positives predicted on the training set
168 true positives are in the training set
The accuracy of true positives on the training set is: 97.619 %
-----------------------------------------------------------------
43 positives predicted on the test set
43 true positives are in the test set
The accuracy of true positives on the test set is: 100.0 %
```

Fig : Accuracy for True positives for Training and Test for malignant tumors

From the above output , it is quite clear that accuracy looks much better on the prediction of the malignant tumors for both training as well as the test set.

```
210 negatives predicted on the training set
211 true negatives are in the training set
The accuracy of true negatives on the training set is: 99.526 %
-----------------------------------------------------------------
138 negatives predicted on the test set
144 true negatives are in the test set
The accuracy of true negatives on the test set is: 95.652 %
```

Fig : Accuracy for True negatives for Training and Test for malignant tumors

This clearly shows that also in the case of the benign tumors , the accuracy looks better on training as well as the test sets.

RESULTS

After running all the machine learning models , it can be interpreted that MLP (Multilayer Perceptron). Since , it is healthcare data and to achieve an optimum benchmark , at least 95% accuracy is needed. Moreover, due to its flexibility it is higher suggested to use it in the real-world scenario.

Model Name	**Training accuracy**	**Test Accuracy**
Radom Forest	99.8%	93.0%
Logistic Regression	89.2%	90.9%
KNN	90.8%	90.2%
Decision Tree	97.2%	90.7%
MLP (Multilayer Perceptron)	98.95%	96.81%

DISCUSSION

Based upon the results it can be clearly said that MLP will used for the better accuracy criteria as it will yield more good results for the purpose of application in Health care industry.

REFERENCES

[1] Breast Cancer Wisconsin (Diagnostic) Data Set. (2020). Retrieved 17 April 2020, from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[2] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization* (Vol. 1905, pp. 861-870). International Society for Optics and Photonics.

[3] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology*, *17*(2), 77-87.

[4] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer letters*, *77*(2-3), 163-171.

[5] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1997). Computerized diagnosis of breast fine-needle aspirates. *The Breast Journal*, *3*(2), 77-80.

[6] Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792-796.

[7] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), 3.

[8] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[9] Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. International Journal of Distributed and Parallel Systems, 4(3), 105.

[10] Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

[11] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31-42.