Cancer Detection using Machine Learning

\*Cancer Diagnosis Detection

Rahul Gupta  
Student, B. Tech., dept of Electronic And Communication Department  
National Institute Of Technology Silchar *Silchar (Assam), India*  
rahulguptaktl123@gmail.com

*Abstract*—Cancer has been distinguished as a composite diseases consisting of many different subtypes. The early diagnosis of a cancer have become very important in cancer research. The necessity of classifying cancer patients into high or low risk group has making many research teams, from the biomedical and bioimfomatics field, to study the application of machine learning . Machine learning is continously increasing in the field of cancer detection ans diagnosis Cancer detection becomes quite easier in future and we can predict it before we suffering from the cancer and before going to the hospitals. As we can see many technologies are being used and tested in this field, so in future this will make us easier to predict cancer. We are testing that which algorithm will us the better result among Random Forest Regression,Decision Tree ,KNN Logistic regression etc.we are going to make a cancer detection model using machine learning and detect the cancer. we are considering are radius, texture, perimeter, area etc. and the result for both cancer is likelihood of being affected by the cancer.

Keywords—Introduction, ML survey in cancer prediction, ML techniques and methods,Results and Discussion,Conclusion.

# Introduction (*Heading 1*)

Over a last few decades , a continous development related to cancer research has been performed. Scientists and Researchers registers different methods, such as screeening in early age ,in order to find the cancer before they cause symptoms.They had expanded new technologies and ideas for the early detection of caner treatment outcomes.With the arrival of new technologies in the field of medicines, large amounts of cancer data has been produced and collected and are available to the medicines research community. However,the presice prediction of diseases outcomes is one of the most interesting and demanding task for the researchers and physicians. As a result ML methods have become well liked tool for medical researchers. These techniques can discover and identfy patterns and relationships between them.

Given the importance of personalized medicines and increasing trends on the applications of ML techniques,we are going to present a review of studies that make use of cancer detection and prognosis. In these studies detection and prognosis features are considered which may be independent of a certain treatement or are desegregated in order to guide therapy for cancer patients. Along with this we are also discussing the types of ML methods being used.

Many studies have been announced in the written works and based on different ideas and strategies that could enable the early cancer diagnosis and detection.Specifically these studies reported the approaches related to the profilling of circulating RNAs that have been proven a promising class for cancer detection and identification. However,these methods suffers from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors.These studies list the potential as well as the limitations of microarray for the prediciton of cancer outocomes. Even gene signatures could rapidly improves our capacity for detection in cancer patients,poor growth had been made for their application in the hospitals.

Problem overview

* 1. The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumor-drug sensitivity) after the diagnosis of the disease. In the latter two situations the success of the prognostic prediction is obviously dependent, in part, on the success or quality of the diagnosis. However a disease prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis

2. ML survey in cancer prediction

ML, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms . There are two main common types of ML

methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme/model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share.

Sometimes conventional statistics proves to be more powerful or more accurate than machine learning. In these casesthe user's initial determinations about the interdependence and nonlinearity of the data would have been wrong. This is not necessarily a weakness to machine learning, it is just a matter of choosing the right tool for the right job. Likewise, not all machine learning methods are created equal. Some are better for certain kinds of problems while others are better for other kinds of problems. For instance some machine learning algorithms scale nicely to the size of the biological domains, others do not. Likewise some methods may have assumptions or data requirements that render them inapplicable to the problem at hand. Knowing which method is best for a given problem is not inherently obvious. This is why it is critically important to try more than one machine learning method on any given training set. Another common misunderstanding about machine learning is that the patterns a machine learning tool finds or the trends it detects are non-obvious or not intrinsically detectable. On the contrary, many patterns or trends could be detected by a human expert – if they looked hard enough at the data. Machine learning simply saves on the time and effort needed to discover the pattern or to develop the classification scheme. Recall that with any interesting discovery, it is frequently obvious to the casual observer – particularly after the discovery has been made.

There are three general types of machine learning algorithms: 1) supervised learning; 2) unsupervised learning and 3) reinforcement learning. They are essentially classified on the basis of desired outcome of the algorithm . In supervised learning algorithms a “prescient provider” or teacher gives the learning algorithm a labeled set of training data or examples. This is the process by which most school children learn. In unsupervised learning, a set of examples are given, but no labels are provided. Instead it is up to the learner to find the pattern or discover the groups. This is somewhat analogous to the process by which most graduate students learn. Unsupervised learning algorithms include such methods as self-organizing feature maps (SOMs), hierarchical clustering and K-means clustering algorithms. These approaches create clusters from raw, unlabeled or unclassified data. These clusters can be used later to develop classification schemes or classifiers.

3. ML techniques and methods

ML algorithms are mainly investigated to decrease computational costs by forecasting the outcome based on the experimental data used to train the ML model. ML has been used in many different research fields, such as biomedicine, signal processing, and the semiconductor industry, due to its viability. Also, complicated processing procedures allow ML to work more carefully in the semiconductor manufacturing industry. In addition, ML uses a variety of statistical models like k-NNs, decision trees, random forests, and linear regression. However, because they maximize the distance between various points, k-NN and linear regression are frequently used for linear dependencies

a. Linear Regression:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is a fundamental technique in the field of statistics and machine learning, particularly in the context of supervised learning.

The basic idea behind linear regression is to find the best-fitting linear equation that describes the relationship between the independent variable(s) and the dependent variable.

b. K Nearest Neighbour Model:

K-Nearest neighbor (KNN) is a supervised learning technique that determines the similarity of a new data point with available cases and puts the new case into the category which is most similar to the available categories.

By choosing K, the number of nearby observations to use in the algorithm is fixed. The algorithm aims to locate all of these K closest neighbours around an unknown new data point and assigns it to the class with the highest closest neighbours. The Euclidean distance between points is calculated to determine the closeness of points.

c. Decision Tree Model:

Decision tree model is a supervised learning technique that uses a decision tree classifier with features of a dataset as the internal nodes, decision rules represented by the branches and the outcome as the leaf nodes. It gives the graphical representation for finding all possible solutions to a decision based on given conditions.

For predicting the class of the given dataset, the algorithm starts from the root node of the tree which contains the entire dataset. This algorithm compares the values of root attribute with the real dataset attribute and, based on the comparison, finds the best attributes in the dataset using the Attribute Selection Measure (ASM) and makes internal nodes. Every internal node that jumps to the subsequent node after the branch and follows it has a decision made for it. The algorithm continues by performing another comparison between the attribute value for the subsequent node and those of the other sub-nodes. This continues until it reaches the leaf node of the tree.

d. Random Forest Regressor Model:

The Random Forest Regressor (RFR) model is a supervised ML algorithm that takes the average of predicted values of multiple decision trees to provide accurate predictions for new data.The input dataset is of m \* n dimensions where m is the total number of data samples and n is size of input features. The RFR model is constructed by creating multiple decision trees in which a subset m’ of the training data (m’ < m) is randomly sampled with replacement (bootstrapped) to create a new dataset for each tree. For each node of the tree, randomly n’ variables (n’ < n) are chosen which forms the basis for the decision at that node. The best split is calculated based on these n’ variables in the training set. Additionally, at each split in the tree, only a random subset of features is considered. This process is repeated until a specified number of trees is grown, resulting in an ensemble of decision trees that are used for predictions. An unbiased average of the resultant predictions from the decision trees is considered to predict output for new input variables. Fig 12 shows the flowchart for the described methodology.

e. Artificial Neural Network

Aritficial neural network is a imformation processing techniques.It’s work like as a human brain process imformation.ANN are biologically inspired computer programs designed to stimulate the way in which human brain does.ANN techniques are used for classification in a neural network.It is a supervised learning techniques.this techniques used to do very complex and non linear data analysis and computation to make give better results and accuracy. In this model we applying the ANN using tensorflow and keras. In this techniques we use relu and sigmoid activation function. The loss is binary crossentropy and using the optimizer ‘adam’ and train the model.

Root Mean Square Error:

The root mean square error (RMSE) function is calculated using the evaluated output by the trained model and the simulated test dataset. For achieving the convergence of error function or to achieve higher accuracy RFR model exhibits ‘low bias’ by populating the trees to its maximum depth and the ‘high variance’ data is converted into ‘low variance’ output by growing each tree on a bootstrap subset of the training dataset and aggregating individual tree predictions to predict the output. The decision trees can be optimised by optimising the hyperparameters such as ‘n\_estimator’ which sets the number of decision trees the model contains, ‘max\_depth’ which sets the maximum depth of a decision tree, ‘min\_samples\_split’ which gives the minimum sample number to split a node and adjusting a few other hyperparameters.

The decision trees are optimized by optimizing the hyperparameters such as ‘no\_of\_estimators’ which sets the number of decision trees the model contains, ‘maxi\_depth’ which sets the maximum depth of a decision tree, ‘mini\_samples\_split’ which gives the minimum sample number to split a node and adjusting a few other hyperparameters.

|  |  |
| --- | --- |
| parameters | **diagnosis** |
| no\_of\_estimators | 30 |
| maxi\_depth | 20 |
| mini\_samples\_split | 2 |
| min\_samples\_per\_leaf | 2 |

Table 1. best parameters for diagnosis of random forest regression.

This model enables us to establish complex non linear relationships between input features and target variables and proves to be a powerful ensemble learning methods.

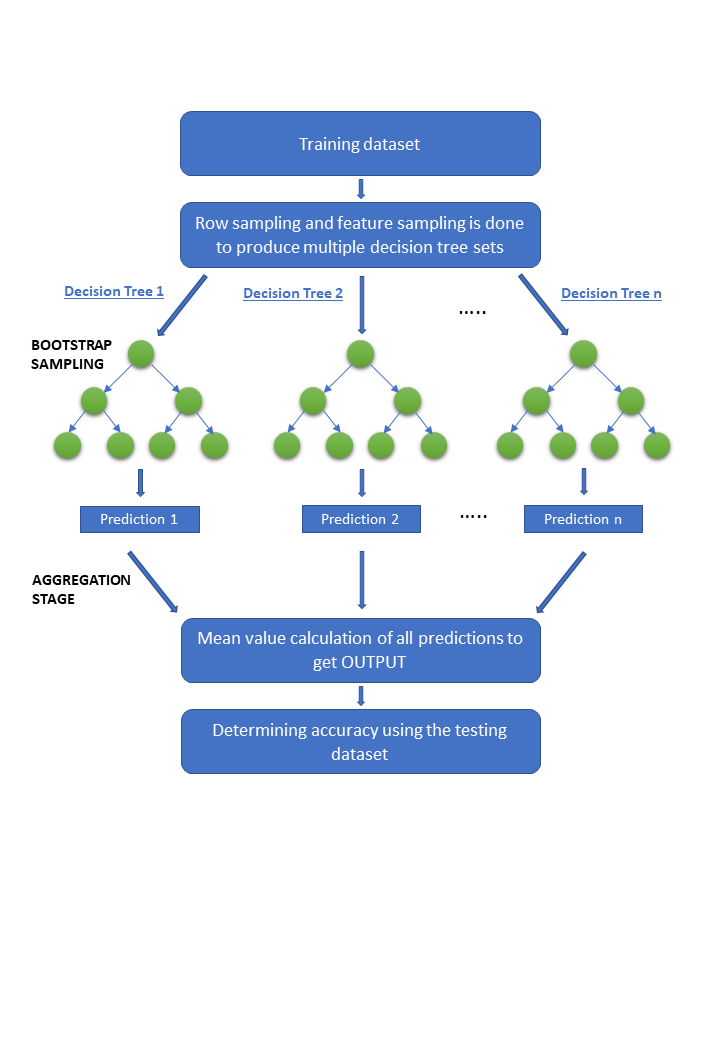


Figure 1: Pictorial representation of Random Forest Regressor (RFR) Model

Some of the advantages of Random forest regressor model are:

1) It is able to extract significant features even from a small dataset

2) The convergence of RFR is faster than the other ML algorithms

3) It is capable of capturing the non linear dependencies as well.

The models proposed in the paper are implemented using the Sklearn library of Python. The dataset obtained is used to train these models and predict future values of diagosis that person has a Malignant and Benign.where malignant means 1 and person has a cancer and benign means 0 and person has not any cancer.

The decision trees are optimized by optimizing the hyperparameters such as ‘no\_of\_estimators’ which sets the number of decision trees the model contains, ‘maxi\_depth’ which sets the maximum depth of a decision tree, ‘mini\_samples\_split’ which gives the minimum sample number to split a node and adjusting a few other hyperparameters.

While training the Machine Learning model, the Root Mean Square Error is calculated. The mathematical formula for calculating the RMSE value for the training or the testing data of the model is given below:

(1)

where N represents the number of samples under consideration and ytrue and ypred represent the actual and predicted values for each of the data point.

Another evaluation parameter considered is the R2 score. The coefficient of determination, or R2, is a measure that provides information about the goodness of fit of a model. It is calculated as given ,

(2)

(3)

The sum of the squares of the data's variance from the mean is the total sum of squares, and the sum squared regression is the sum of the squared residuals. Since it is a percentage, it can only accept values in the range of 0 and 1.

4. Results And Discussions

In the above section we have discussed about the cancer diagnosis and how we know the person has cancer or not from the variou input ideal range. The values of the hyper parameters for the Random Forest Regression algorithm under testing are shown in table 1. Table 2 below shows the best RMSE values obtained for various models for each of the output parameters. The RMSE error values of the RF model are lesser than the errors for other models.

|  |  |  |
| --- | --- | --- |
| Training Algorithms | SCORE | RMSE |
| Linear Regression | .71 | 0.26 |
| Logistic Regression | .94 | 0.22 |
| K Nearest Neighbour Regression | .93 | 0.247 |
| Decision Tree Regression | .92 | 0.28.098 |
| Random Forest Regressor | .87 | 0.17 |
| Artificial Neural Network | .98 | .098 |

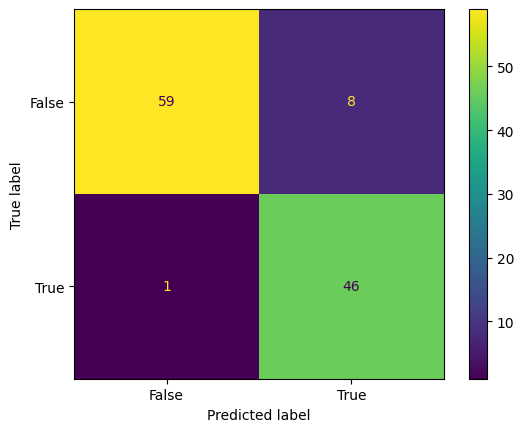
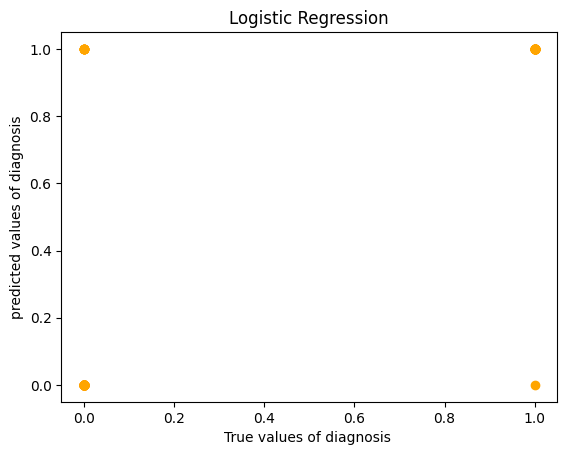
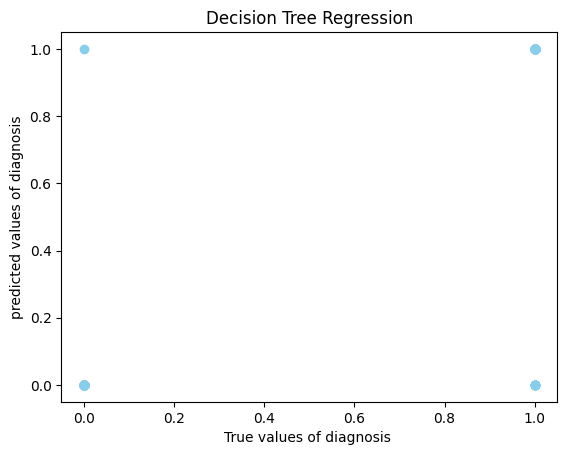
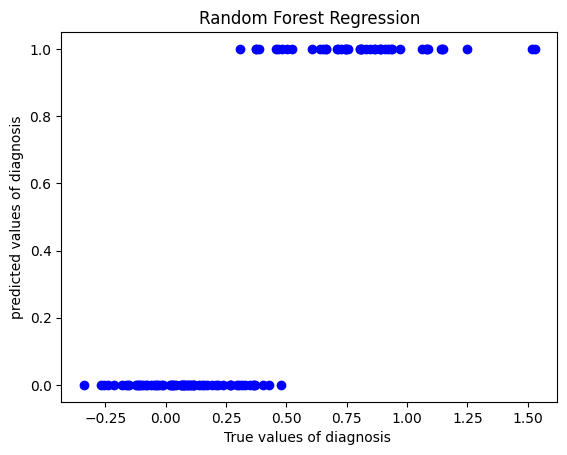
Fig2. Diagram of Confusion Matrix for the predicted values of diagnosis and True values.

Fig4. shows the confusion matrix which is plot between the true values and predicted values this matrix showthat how much our model’s predicted values are accurate according to the true values and matrix also tell the more the darker colour implies low prediction values lies there and more the lighter colour is showing the higher the acurate prediction.For first box tell the there is a only one value for which truth label is true and predicted label is false and second box say the there are 46 values for which truth label and predicted label is true, and third box say that there are 59 values for which there is truth and predicted labels are false mean when the patient has no cancer,our model also predicted that same result as we have.







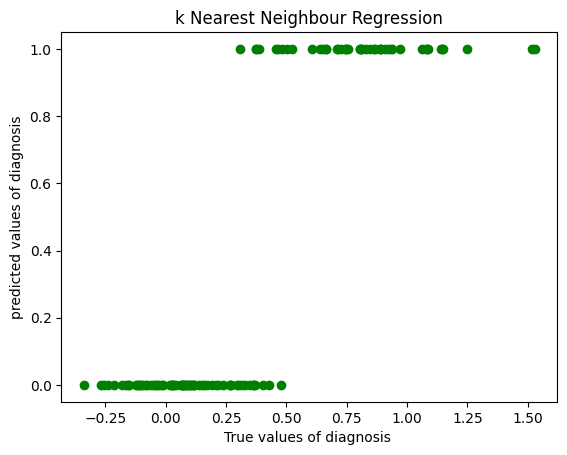
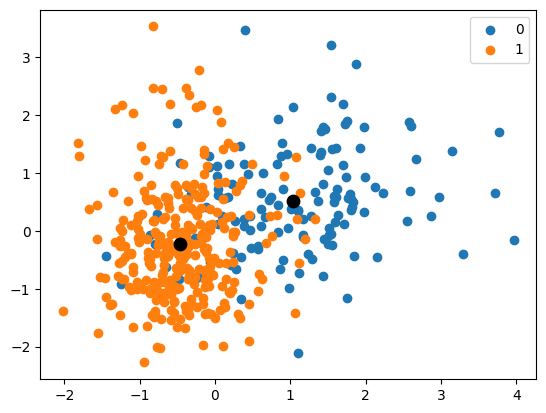
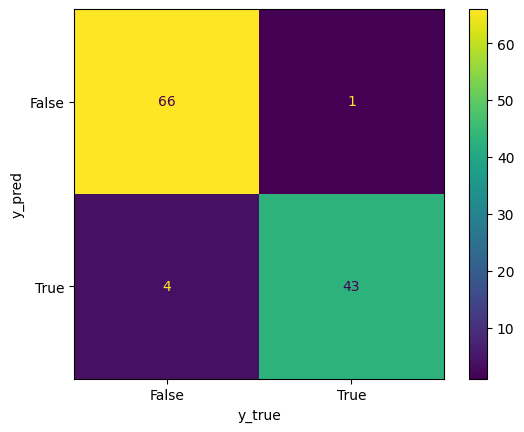


Fig3. Scatter plot of predicted value and true value of diagnosis.

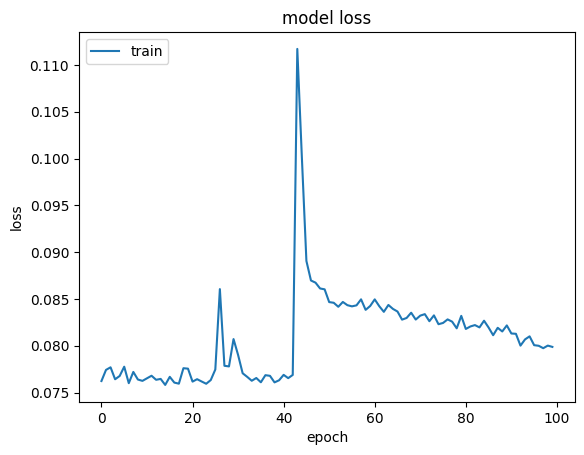
The above given figures are scatter plot for the differnet model of regression and classification techniques between the true values and predicted values of diagnosis.First one is scatter plot of linear regression and we easily see that the output has only two values 0’s and 1’s.Second one is the scatter plot between true values and predicted values logistic regression is a classification technique so we get this type of scatter plot in which we can see only 0 and 1 values at four points.Third one is Decision Tree Regression which is also a classification technique and we get the same type of scatter plot as same as logistic regression.Fourth is Random Forest Regression in which we get the scatter plot as same of linear regression.Last one is knn which is also a classfication technique

fig 4. clustering of range of diagnosis values.

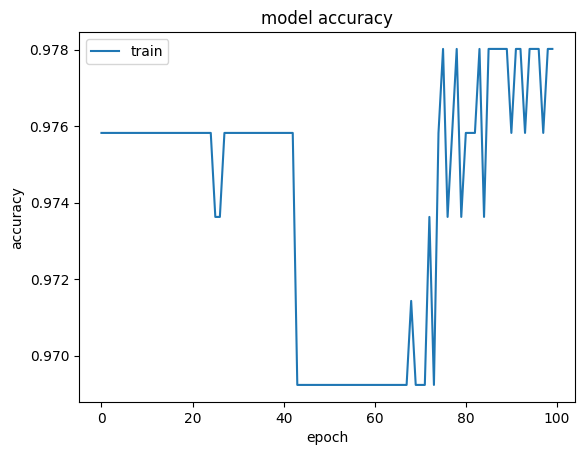
The above figure show the clustering of two ranges of output values 0 and 1.These are the values on graph that the people has cancer or not in which blue show that patient has not any cancer and orange show that the patient has cancer.The two black points between the clusters show the center of both the cluster .



classfication report of ANN model



model loss of ANN with respect to epochs.



model accuracy with respect to epochs.

5.Conclusion

The Model that we have discussed is about cancer detection that person has a cancer which is represent by malignment and person hasn’t any cancer which is represent by benign. From this model.The combinantio of both diagnosis and machine learning result used in triaining and prediction.The five regression techniques like Linear Regression,Logistic Regression,Decision Tree forest,Random Forest Regression and K Nearest Neifgbour Regression are explored to observe their ability of prediction with the given datasets.The perfomance of these algorithms are evualated by using the root mean squared error(RMSE) and r2 score as perfomance parameter.The Logistic Regression algorithm show the better prediction accuracy with 94% score.

This work can be extended by generating a few more data and exploring the new deep learning techniques to perform the prediction of output.

END

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**