International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cancer Prediction Using Machine Learning

Nadipalli Shravani

Student, Department of Information Technology, GMR Institute of Technology, Rajam, India

ABSTRACTION

Machine learning is increasingly being employed in cancer prediction. Cancer prediction will become quite easy within the long run which we are able to predict it without the requirement of visiting the hospitals. As we are going to see many technologies have gotten used and tested within the medical field. So, by this we'll say that this will make us easier within the long run to detect cancer. We are testing which algorithm will give us good result among SVM. We are making a cancer prediction using machine learning, during which we are including three kinds of cancer they're carcinoma, lungs cancer and carcinoma. In carcinoma, we are using SVM algorithm and for lung and prostate we are using Random Forest, Decision tree algorithm. We are visiting give different attributes for 3 cancer system where the user has to enter data to urge result. For carcinoma we are considering attributes like clump thickness, uniform cell size, uniform cell shape etc. and so the prediction result are visiting be whether the cancer is malignant or benign. For carcinoma, we are considering smoking, yellow fingers, anxiety, peer pressure etc. In carcinoma, we are considering are radius, texture, perimeter, area etc. and thus the result for both cancer is likelihood of being littered with the cancer. In this paper we will discuss these algorithms and apply those algorithms to data sets to predict the cancer.

Keywords: Machine Learning, Decision tree, Random Forest, svm,

INTRODUCTION

Cancer is a complaint claims hundreds of lives, every time. utmost of the people knows atleast one person who suffers from cancer. In fact, its such a wide affection that it's the alternate loftiest leading because of death in the us, right behind the heart complaint. So it should come as no surprise that numerous of those in the medical field have made it their primary focus for exploration. One of the biggest areas of exploration when it comes to cancer prediction. Because veritably soon cancer is detected the advanced chance that the case will survive. In utmost cases, cancer is diagnosed after it has formerly entered into the advanced stages, which oppressively decreases the effectiveness of current cancer treatments. Trying to prognosticate cancer int its earlier stages, or indeed before it happens, gives us a better treatment outgrowth for the cases.

LITERATURE SURVEY:

1.C. Arunkumar,S. Ramakrishnan, Prediction of cancer using customised fuzzy rough machine learning approaches, Healthcare Technol.Lett., 6(1)(2019),pp. 13- 18In particular, we looked at the frequence with which " cancer vaticination prognostic styles " and " cancer threat assessment vaticination styles " passed in PubMed. These queries yielded 1061 and 157 successes independently, giving anon-overlapping set of 1174 papers. Removing the 53 papers with machine literacy factors in this set, we were left with 1121 papers. While a detailed review of each epitome wasn't possible, a arbitrary slice indicated that

80 of these papers were applicable(890 papers) in that they used statistical approaches to prognosticate or predict cancer issues. thus these data suggest that machine literacy styles regard for103/890(11) of all PubMed papers describing cancer vaticination or prognostic methodology. Overall, the same monthly growth trends(i.e. near exponential) in vaticination and prognostic were observed for the statistical styles as for the machine learningmethods.When looking at the types of prognostications or vaticinations being made, the vast maturity(86) are associated with prognosticating cancer mortality(44) and cancer rush(42). still, a growing number of more recent studies are now aimed at prognosticating the circumstance of cancer or the threat factors associated with developing cancer. As a general rule, anyhow of the machine literacy system used, the type of vaticination being made or the type of cancer being estimated, machine literacy styles appear to ameliorate the delicacy of prognostications by and normal of 15 – 25 over indispensable or conventional approaches

2.J. Pati, Gene expression analysis for early lung cancer vaticination using machine literacy ways aneco-genomics approach, IEEE Access, 7(2019), pp.

Of the 79 papers surveyed in this review, fairly many papers(just 3) employed machine literacy to prognosticate cancer threat vulnerability. One of the more intriguing papers(Listgarten etal. 2004), used single nucleotide polymorphism(SNP) biographies of steroid metabolizing enzymes(CYP450s) to develop a system to retrospectively prognosticate the circumstance of " robotic " bone cancer. robotic ornon-familial bone cancer accounts for about 90 of all bone cancers(Dumitrescu and Cotarla 2005). The thesis in this study was that certain combinations of steroid- metabolism gene SNPs would lead to the increased accumulation of environmental poisons or hormones in bone towel leading to a advanced threat for bone cancer. The authors collected

SNP data(98 SNPs from 45 different cancer- associated genes) for 63 cases with bone cancer and 74 cases without bone cancer(control). crucial to the success of this study was the fact that the authors employed several styles to reduce the sample- per- point rate and delved multiple machine literacy styles to find an optimal classifier. Specifically, from a starting set of 98 SNPs the authors snappily reduced this set to just 2 – 3 SNPs that sounded maximally instructional. This reduced the sample- per- point rate to a respectable 451(for 3 SNPs) and 681(for 2 SNPs) rather of close to 32(had all 98 SNPs been used). This allowed the study to avoid falling victim to the " curse of dimensionality "(Bellman 1961; Somorjai etal. 2003). Once the sample size was reduced, several machine literacy ways were employed including a naïve Bayes model, several decision tree models and a sophisticated support vector machine(SVM). The SVM and naïve Bayes classifiers attained the loftiest delicacy using only a set of 3 SNPs and the decision tree classifier attained the loftiest delicacy using a set of 2 SNPs. The SVM classifier performed the stylish with an delicacy of 69, while the naïve Bayes and decision tree classifiers achieved rigor of 67 and 68, independently. These results are roughly 23 – 25 better than chance. Another notable point to this study was the expansive position of cross confirmation and evidence performed. The prophetic power of each model was validated in at least three ways. originally, the training of the models were assessed and covered with20-foldcross-validation

3.M. Agrawal," Cancer Prediction Using Machine Learning Algorithms", International Journal of Science and Research, 2018

Nearly half of all machine literacy studies on cancer vaticination were concentrated on prognosticating patient survivability(either 1 time or 5 time survival rates). One paper of particular interest(Futschik etal. 2003) used a mongrel machine literacy approach to prognosticate issues for cases with verbose large B- cell carcinoma(DLBCL). Specifically, both clinical and genomic(microarray) data were combined to produce a single classifier to prognosticate survival of DLBCL cases. This approach differs kindly from the study of Listgarten etal.(2004) which only employed genomic(SNP) data in its classifier schema. Futschik etal. hypothecated, rightly, that clinical information could enrich microarray data similar that a combined predictor would perform better than a classifier grounded on either microarray data alone or clinical data alone. In assembling the test and training samples, the authors collected microarray expression data and clinical information for 56 DLBCL cases. The clinical information was attained from the International Prediction Index(IPI) which consists of a set of threat factors, that when duly assessed, allows cases to be separated into groups ranging from low- threat to high- threat. The data from the case's IPI groups was also used to produce a simple Bayesian classifier. This classifier achieved an delicacy of73.2 in prognosticating the mortality of DLBCL cases. Independently from the Bayesian classifier, several different types of " evolving fuzzy neural network "(EFuNN) classifiers were also developed to handle the genomic data. The stylish EFuNN classifier used a subset of 17 genes from the microarray data. This optimal EFuNN had an delicacy of78.5. The EFuNN classifier and the Bayesian classifier were also combined into a hierarchical modular system to induce a agreement vaticination. This mongrel classifier attained an delicacy of87.5, a clear enhancement over the performance of either classifier alone. This was also 10 better than the stylish performing machine literacy classifier(77.6 by SVMs).

4. N. Picco, R.A. Gatenby, A.R.A. Anderson, Stem cell plasticity and niche dynamics in cancer progression, IEEE Trans. Biomed. Eng., 64 (3) (2017), pp. 528-537

A total of 43% of the studies analyzed for this review applied machine learning towards the prediction of cancer relapse or recurrence. One particularly good example is the study of De Laurentiis et al. (1999), which actually addresses some of the drawbacks noted in the previous studies. These authors aimed to predict the probability of relapse over a 5 years period for breast cancer patients. A combination of 7 prognostic variables was used including clinical data such as patient age, tumor size, and number of axillary metastases. Protein biomarker information such as estrogen and progesterone receptor levels was also included. The aim of the study was to develop an automatic, quantitative prognostic method that was more reliable than the classical tumor-node-metastasis (TNM) staging system. TNM is a physician-based expert system that relies heavily on the subjective opinion of a pathologist or expert clinician. The authors employed an ANN-based model that used data from 2441 breast cancer patients (times 7 data points each) yielding a data set with more than 17,000 data points. This allowed the authors to maintain a sample-to-feature ratio of well over the suggested minimum of 5 (Somorjai et al. 2003). The entire data set was partitioned into three equal groups: training (1/3), monitoring (1/3), and test sets (1/3) for optimization and validation. In addition, the authors also obtained a separate set of 310 breast cancer patient samples from a different institution, for external validation. This allowed the authors to assess the generalizability of their model outside their institution—a process not done by the two previously discussed studies.

4.N. Picco,R.A. Gatenby,A.R.A. Anderson, Stem cell malleability and niche dynamics in cancer progression, IEEE Trans. Biomed.Eng., 64(3)(2017),pp. 528- 537

A aggregate of 43 of the studies anatomized for this review applied machine literacy towards the vaticination of cancer relapse or rush. One particularly good illustration is the study of De Laurentiis etal.(1999), which actually addresses some of the downsides noted in the former studies. These authors aimed to prognosticate the probability of relapse over a 5 times period for bone cancer cases. A combination of 7 prognostic variables was used including clinical data similar as patient age, excrescence size, and number of axillary metastases. Protein biomarker information similar as estrogen and progesterone receptor situations was also included. The end of the study was to develop an automatic, quantitative prognostic system that was more dependable than the classical excrescence- knot- metastasis(TNM) staging system. TNM is a croaker - grounded expert system that relies heavily on the private opinion of a pathologist or expert clinician. The authors employed an ANN- grounded model that used data from 2441 bone cancer cases(times 7 data points all) yielding a data set with further than,000 data points. This allowed the authors to maintain a sample- to- point rate of well over the suggested minimum of 5(Somorjai etal. 2003). The entire data set was partitioned into three equal groups training(1/3), monitoring(1/3), and test sets(1/3) for optimization and confirmation. In addition, the authors also attained a separate set of 310 bone cancer case samples from a different institution, for external confirmation. This allowed the authors to assess the generalizability of their model outside their institution — a process not done by the two preliminarily bandied studies.

5..L. Demidova,I. Klyueva,Y. Sokolova,N. Stepanov,N. Tyart, Intellectual approaches to enhancement of the bracket opinions quality on the base of the SVM classifier, Procedia Comput.Sci., 103(2017),pp. 222- 230

Nearly half of all machine literacy studies on cancer vaticination were concentrated on prognosticating patient survivability(either 1 time or 5 time survival rates). One paper of particular interest(Futschik etal. 2003) used a mongrel machine literacy approach to prognosticate issues for cases with verbose large B- cell carcinoma(DLBCL). Specifically, both clinical and genomic(microarray) data were combined to produce a single classifier to prognosticate survival of DLBCL cases. This approach differs kindly from the study of Listgarten etal.(2004) which only employed genomic(SNP) data in its classifier schema. Futschik etal. hypothecated, rightly, that clinical information could enrich microarray data similar that a combined predictor would perform better than a classifier grounded on either microarray data alone or clinical data alone. In assembling the test and training samples, the authors collected microarray expression data and clinical information for 56 DLBCL cases. The clinical information was attained from the International Prediction Index(IPI) which consists of a set of threat factors, that when duly assessed, allows cases to be separated into groups ranging from low- threat to high- threat. The data from the case's IPI groups was also used to produce a simple Bayesian classifier. This classifier achieved an delicacy of73.2 in prognosticating the mortality of DLBCL cases. Independently from the Bayesian classifier, several different types of " evolving fuzzy neural network "(EFuNN) classifiers were also developed to handle the genomic data. The stylish EFuNN classifier used a subset of 17 genes from the microarray data. This optimal EFuNN had an delicacy of78.5. The EFuNN classifier and the Bayesian classifier were also combined into a hierarchical modular system to induce a agreement vaticination. This mongrel classifier attained an delicacy of87.5, a clear enhancement over the performance of either classifier alone. This was also 10 better than the stylish performing machine literacy classifier(77.6 by SVMs).

Introduction to Machine Learning:

Machine learning could be a subfield of artificial intelligence(AI) . The goal of machine learning typically is to grasp the structure information|of knowledge|of information} and work that data into modules which will be understood and used by people.

Although machine learning could be a field inside technology , it differs from ancient process approaches. In ancient computing ,algorithm are sets of expressly programmed directions utilized by computers to calculate or downside solve. Machine learning algorithms instead afford computers to coach on knowledge inputs and use applied mathematics analysis facilities computers in building models from sample knowledge so as to alter decision-making processes supported knowledge inputs

Machine Learning could be a continuous developing field. attributable to this, there are some concerns to stay in mind as you're employed with machine learning methodologies, or analyze the impact of machine learning method.

Machine Learning could be a term closely related to knowledge science . It refers to a broad category of ways that revolves around knowledge modelling to:

(1) ***Supervised Learning*** : to algorithmically build predictions, and

(2) ***Unsupervised Learning***: to algorithmically decipher patterns in knowledge

Supervised Machine Learning:

It is used for structures datasets. It analyses the coaching knowledge and generates operations which is able to be used for different datasets. it's Machine Learning for creating predictions-Core idea is to use labeled knowledge to coach prognostic models. coaching models suggests that mechanically characterizing labelled knowledge in ways that to predict tags for unknown knowledge points. for instance a master card fraud detection model may be trained employing an account of labelled fraud purchases. The resultant model estimates the chance that any new purchase is deceitful. Common ways for coaching models vary from basic regressions to advanced neural nets. All follow identical paradigm grasp as supervised learning

3.1 Machine learning techniques used:

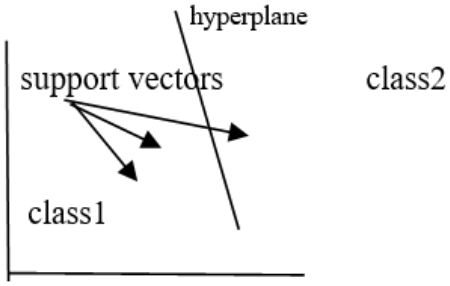
Decision Tress: Simply put a decision tree in which each branch node represents a choice between no of alternatives and each nodes represents a decision .

It is a type of supervised learning algorithm(predefined target value) that is mostly used in classification problems works on the both categorical and continuous input and output variable it is a one of the widely used and pratical methods for inductive inference (inductive inference is the process of reaching general conclusion from specific examples)

Decision tree learn and trained themselves from given example and predict for unseen circumstances .The following is the block diagram of decision tree learning

Block Diagram of Decision Tress

Support Vector Classifier-. SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges.However,it is mostly used for classification problems.In this algorithm,we plot each data items as a point n-dimenstional space(where n is number of features you have)with the value of each features being the value of a particular coordinate. Then,we perform classification by finding the hyper-plane(in two dimentional space it is classifier line, mostly strait)that differentiate the two classes (groups of data)very well



Supper Vector are merely the co-ordinates of individual observation. Support Vector Machine could be a frontier that best segregates the 2 categories (hyper-plane/line) the method of segregation the 2 categories with hyperplane. Support vector machine (SVM) once your information has precisely 2 categories. associate SVM categories information by finding the simplest hyperplane that separates all information points of 1 category from those of the opposite class.The best hyperplane for associate SVM means the one with the massive margin between the 2 categories. Margin means that the largest breadth of the block parallel to the hyperplane that has no interior information points

Random Forest – is Associate in Nursing extension over sacking. It takes one further step wher additionally to taking the random set of information,it conjointly takes the random choice of options instead of victimization all options to grow hair style. after you have several random trees. It's known as Random Forest.

Steps taken to implement Random Forest:

1. 1.Suppose there square measure N observations and M options in coaching knowledge set.First , a sample from coaching knowledge set is taken every which way with replacement
2. A set of M options square measure designated every which way and whichever feature offers the simplest split is employed to separate the node iteratively
3. The tree is big to the biggest
4. Above steps square measure perennial and prediction is given supported the aggregation of predictions from n variety of trees

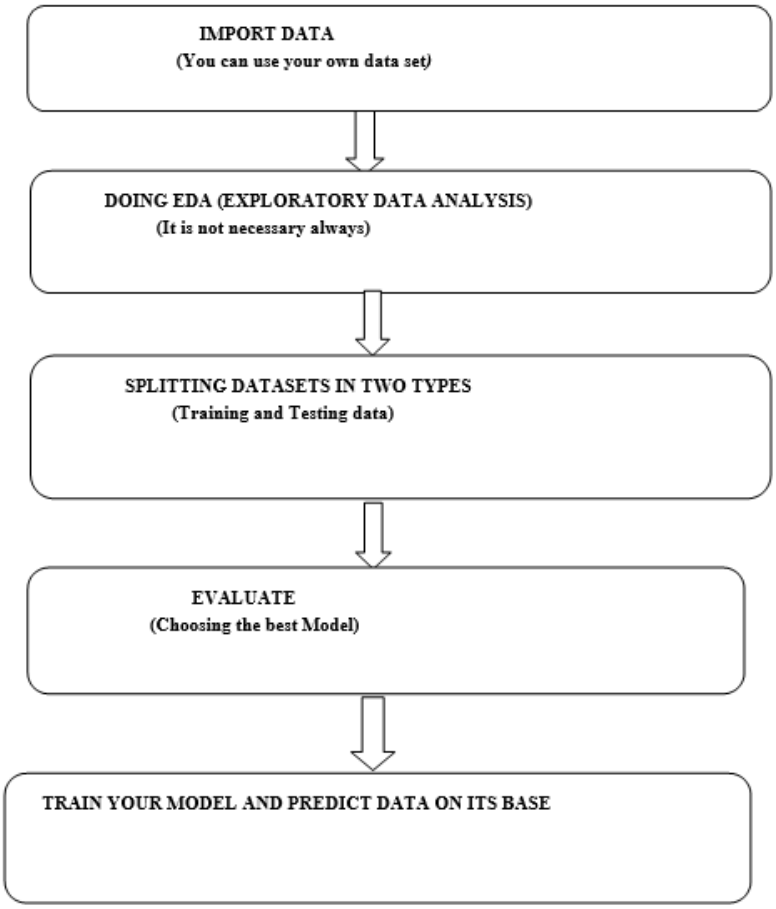
Random Forest may be a versatile machine-learning methodology capable of playacting each regression and classification tasks. It conjointly undertakes dimensional reduction ways, treats missing values, outliers values, and different essential steps of information exploration. it's a kind of ensemble learning methodology, wherever a bunch of weak modules combines to make a strong model

In Random Forest, we have a tendency to grow multiple trees as opposition one tree in CART model. To classify a replacement object supported Associate in Nursing attribute, every tree offers a categoryification and that we say the tree "votes" for that class. The forest chooses the classification having the foremost votes(over all the trees within the forest) and just in case of regression, it takes the common of outputs by completely different hair styleDesigning Machine Learning Algorithms***:***

The following are the ingradients of style procedure:

1. ***check Harnessing***: you wish to outline a check hardness. The check harness is that the knowledge you'll train associated check an algorithmic rule against and therefore the performance live you'll use to assess its performance. it's necessary to outline your check harness well so you'll concentrate on evaluating completely different algorithms and thinking deeply concerning the matter
2. ***Performance Measure:*** The performance live is that the means you would like to judge an answer to the matter. it's the measuring you'll create of the predictions created by a trained model on the check knowledge set
3. ***Testing and coaching Datasets:*** From the remodeled knowledge, you'll have to be compelled to choose a check set and a coaching set. associate algorithmic rule are trained on the coaching dataset and can be evaluated against the check set. this might be as straightforward as choosing a random split of data(70% for coaching,30% for testing) or could involve a lot of difficult sampling strategies**.**
4. ***Cross Validation:*** A lot of subtle approach than employing a check and train knowledge set is to use the whole remodeled knowledge set to coach and check a given algorithmic rule. a way you'll use in your check harness that will this is often known as cross-validation

FLOW CHART:



PROPOSED WORK

This proposed system presents a comparison of machinelearning (ML) algorithms: Support machine vector(SVM),Decision Tree(DT), Random Forest(RT), Artificial Neural Net- works(ANN), Naive Bayes (NB), Nearest Neighbour (NN) search. The data-set used is obtained from the Wisconsin datasets.For the implementation of the ML algorithms, the dataset was partitioned into the training setand testing set. A comparison between all the six algorithms will be made. The algorithm that provides the best results will be supplied as a model to the web site . the web site will bemade from a python framework, called flask. And it'll host the database on Xampp or Firebase or inbuilt Python and flask libraries. This data set is out there on the UCI Machine Learning Repository. It consists of 32 world attributes which are multivariate. the entire number of nstances is 569 and there are not any missing values in thisdata set. the method of the proposed system is as follows,

1. The patient books a meeting through our website.
2. The patient will then meet the doctor offline for the respective appointment.
3. The doctor will first check the patient manually, then perform a breast mammogram or an ultrasound. That ultra sound will show a picture of the breast consisting the lumps or not.
4. If the lumps are detected, a biopsy are going to be performed. The digitised image of the Fine Needle Aspirate (FNA) is what forms the features of the dataset.
5. Those numbers are going to be provided to the system by the doctor and therefore the model will detect if it’s a benign or a malignant cancer.
6. The report are going to be then forwarded to the patient on their respective account

Regression Metrics:

The following three are the most common metrics for evaluating predictions on regression machine learning problems:

1. ***Mean Absolute Error:*** is the sum of the absolute difference between predictions and actual values. It gives an idea of how wrong the predictions were. The measure gives an idea of the magnitude of the error, but no idea of direction(e.g. over or under-predicting).

2. ***Mean Squared Error(MSE)***:MSE is much like the mean absolute error in that it provides a gross idea of the magnitude of error. It is a measure of the difference between two continuous variables. Assume X and Y are variables of paired observations that express the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. Consider a scatter plot of n points, where point I has coordinates .Mean Absolute Error(MSE) is the average vertical distance between each point and the Y=X line, Which is also known as the One-to-one line MAE is also the average horizontal distance between each point and the line Y=X line

3. ***R-Squared Metric***:the R-Squared metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determine. This value between 0 and 1 for no-fit and perfect fit respectively

CONCLUSION

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes.

REFERENCES:

1. C. Arunkumar , S. Ramakrishnan, Prediction of cancer using customised fuzzy rough machine learning approaches, Healthcare Technol. Lett., 6 (1) (2019), pp. 13-18

2. J. Pati, Gene expression analysis for early lung cancer prediction using machine learning techniques: an eco-genomics approach, IEEE Access, 7 (2019), pp.

3. M. Agrawal, "Cancer Prediction Using Machine Learning Algorithms", International Journal of Science and Research, 2018

4. N. Picco, R.A. Gatenby, A.R.A. Anderson, Stem cell plasticity and niche dynamics in cancer progression, IEEE Trans. Biomed. Eng., 64 (3) (2017), pp. 528-537

5. L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart, Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier, Procedia Comput. Sci., 103 (2017), pp. 222-230